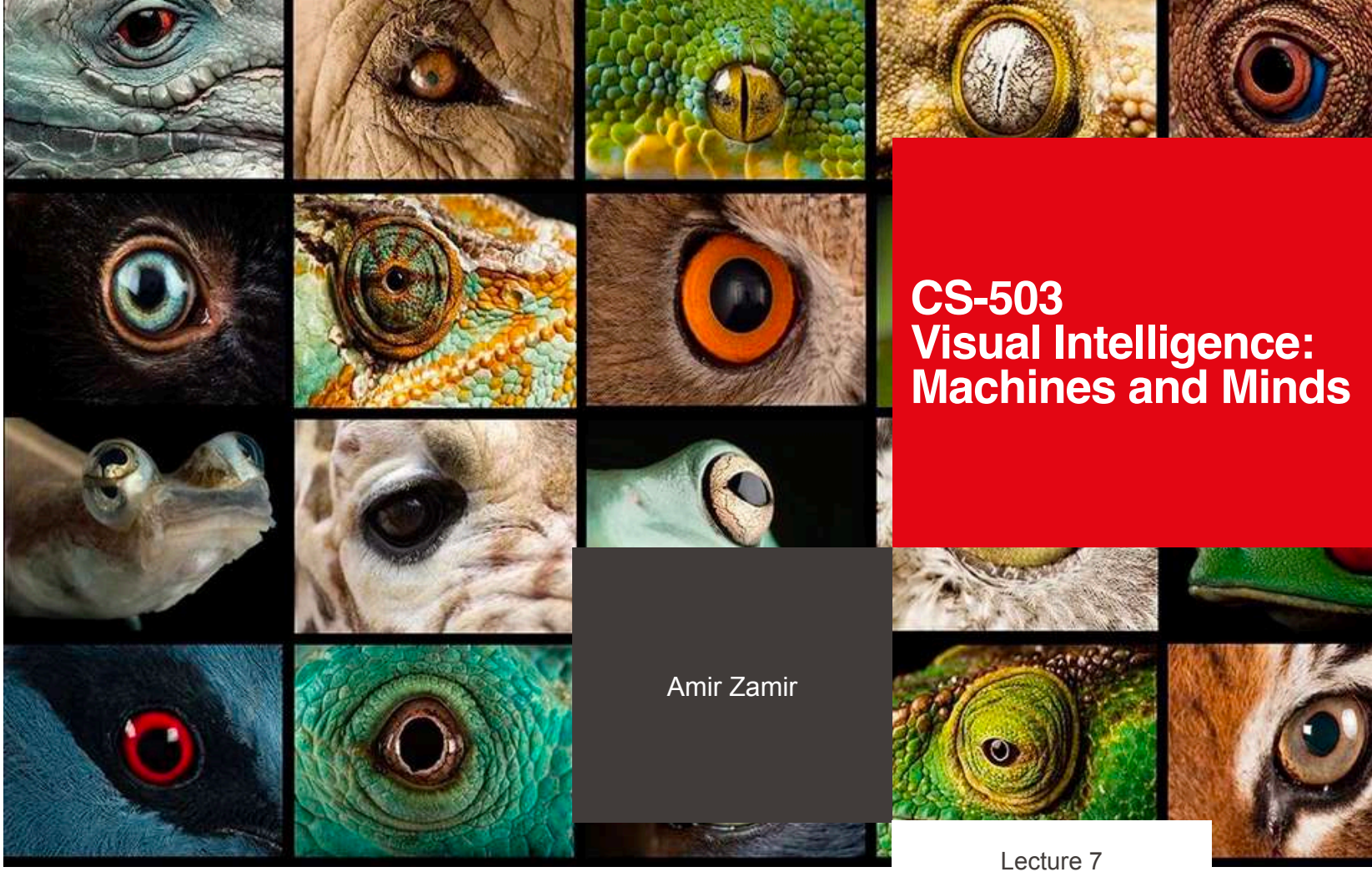


CS-503
Visual Intelligence: Machines and Minds

Amir Zamir

Lecture 7

École
polytechnique
fédérale
de Lausanne

Logistics

Logistics

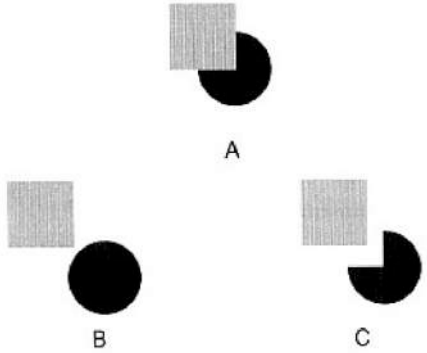
- Next assignment notebook due 01/04/2025 23:59 CET.

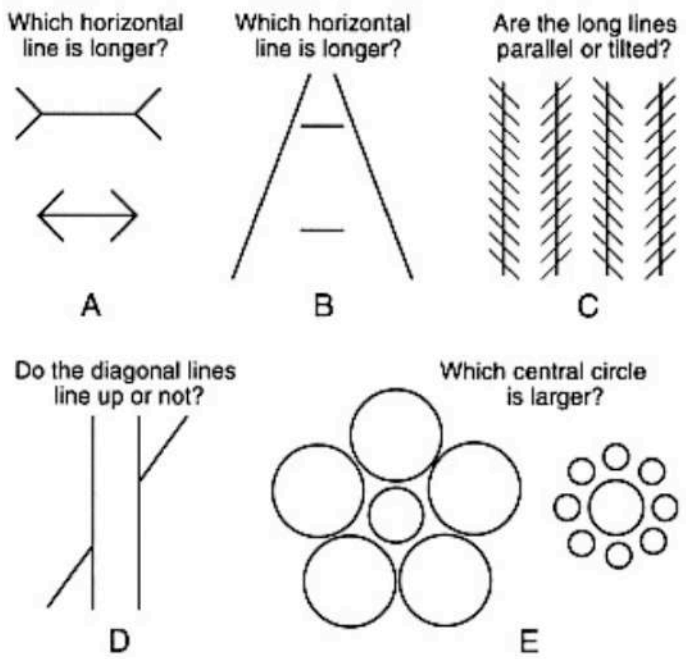
Week Num.	Date	Item
1	20.02	- lecture 1
2a	25.02	- lecture 2
2b	27.02	- lecture 3
3a	04.03	- lecture 4
3b	06.03	- lecture 5
4a	11.03	- lecture 6 (+ Q&A)
	11.03	- Transformers notebook assignment due
4b	13.03	- lecture 7
5a	18.03	- lecture 8
5b	20.03	- lecture 9
6a	25.03	- lecture 10
6b	27.03	- lecture 11 (+ Q&A)
	01.04	- Active agents notebook assignment due
7a	01.04	- lecture 12
7b	03.04	- lecture 13
8a	08.04	- lecture 14
8b	10.04	- lecture 15 (+ Matchmaking session)
	13.04	- Project proposals due
	15.04	- all subsequent sessions from 15.04 onwards are for Q&A
	18.04	- Project proposals due, when revision is needed.
	22.04	- MidSem break - No classes
	25.04	- MidSem break - No classes
	29.04	- Foundation Models assignment due
	01.05	- lecture 16
	09.05	- Project progress report due
	13.05	- Robustness assignment due (extra credit)
	20.05	- Moodle homework due
	26.05	- Final project presentation video due
	27.05	- Final project presentation Part I
	29.05	- Final project presentation Part II
	30.05	- Project report due

Recap

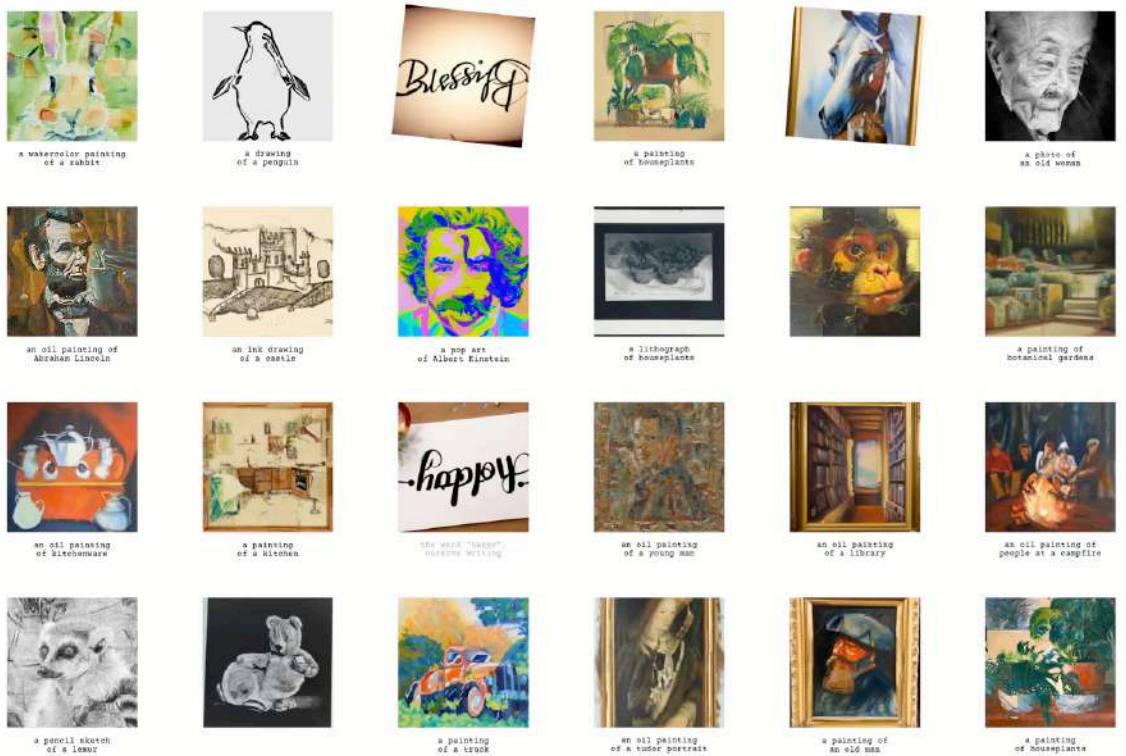
Constructivism

- Vision:
  - an indeterminate inverse problem from retinal images.
  - a “reconstruction” of the reality.
- Something besides the retinal image is needed.
- Likelihood Principle



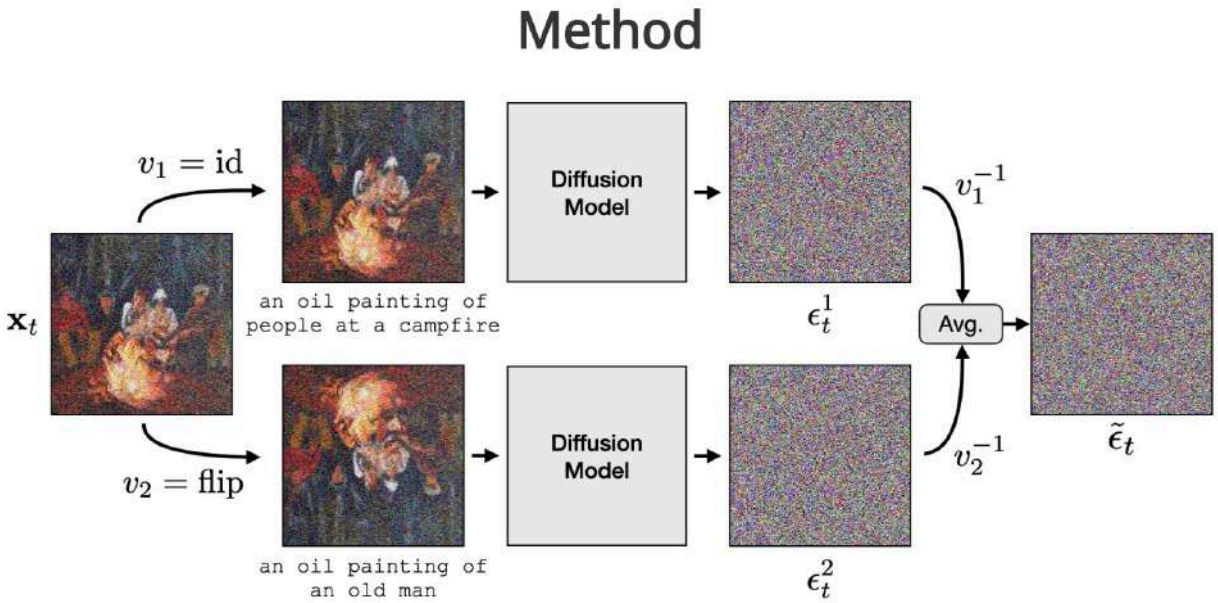


Generated optical illusions



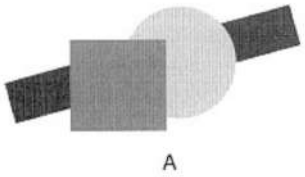
Visual Anagrams: Generating Multi-View Optical Illusions with Diffusion Models, Geng et al., CVPR 2024 (Oral)
https://dangeng.github.io/visual_anagrams/

Generated optical illusions

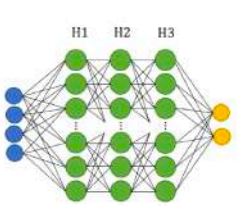


Visual Anagrams: Generating Multi-View Optical Illusions with Diffusion Models, Geng et al., CVPR 2024 (Oral)
https://dangeng.github.io/visual_anagrams/

Perception as modeling the environment

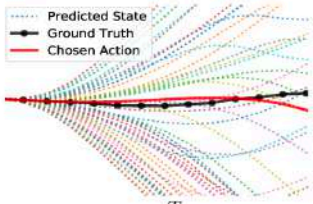


- How and why?
  - The evolutionary utility of vision toward survival and reproduction, in the environment.
- *The observer is constructing a **model** of what environment situation might have produced the observed pattern of sensory stimulation*



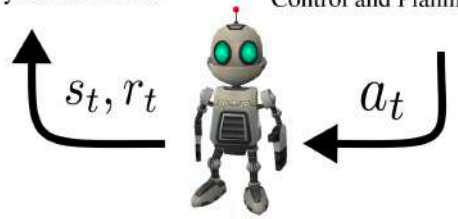
$s_{t+1} = f_\theta(s_t, a_t)$

Dynamics Model

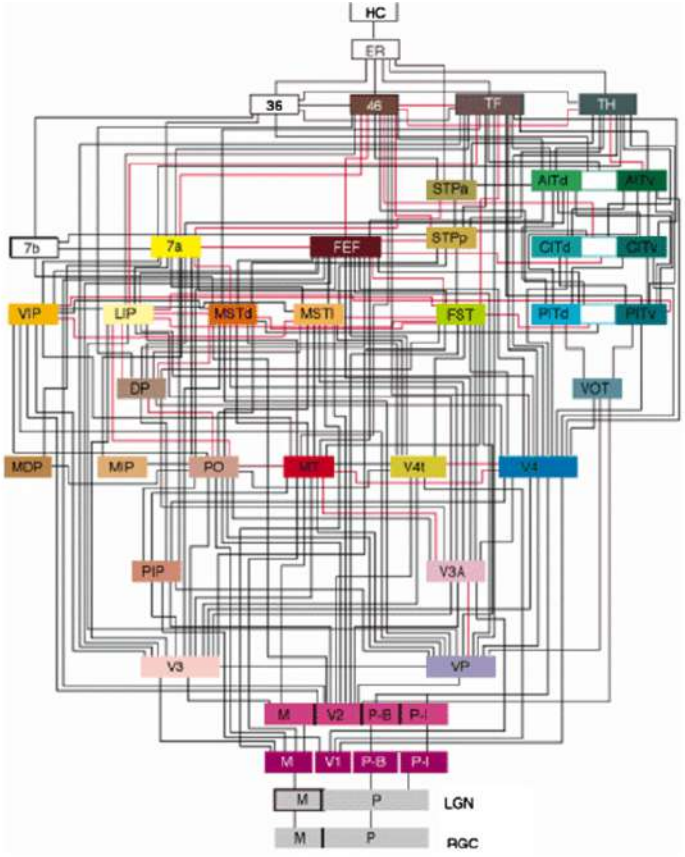


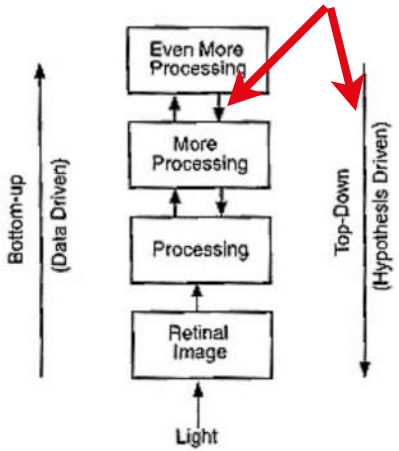
$\arg\max_{u(t)} \sum_{t=0}^{T} r(s_t, a_t)$

Control and Planning



Feedback

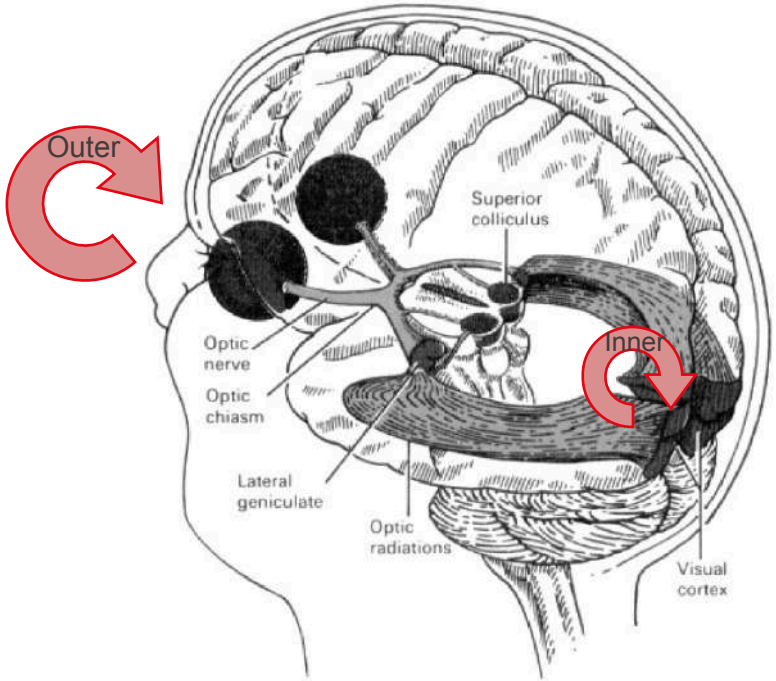




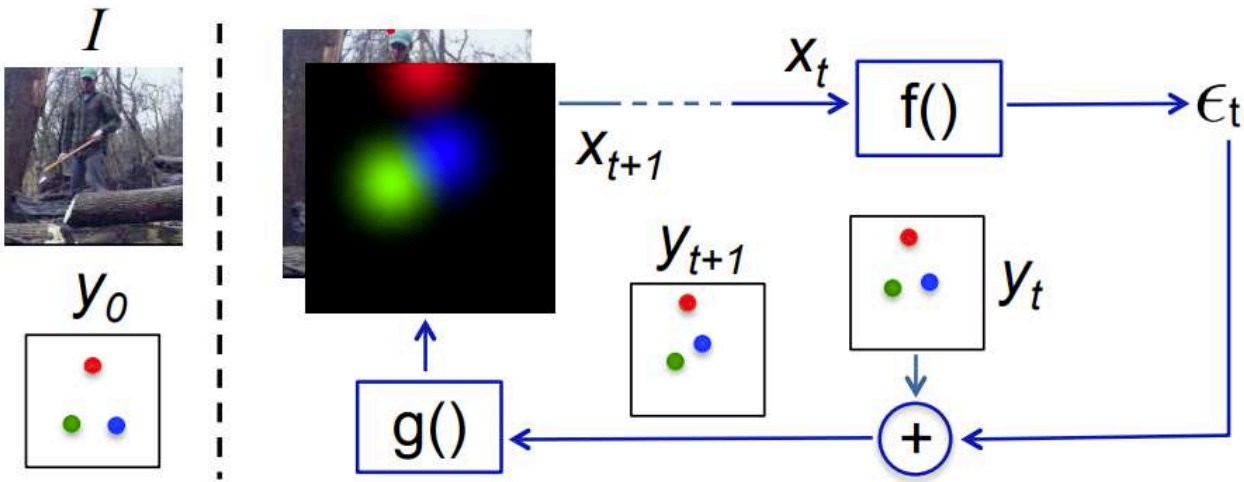
- Felleman & Van Essen (1991)

Feedback

- Inner loop
  - top-down processing without external feedback from the world.
  - e.g. IEF (iterative error feedback, 2016), Attention, Feedback Networks (2017), diffusion.
- Outer loop
  - with external feedback from the world
  - e.g. RMA (2021), RNA (2023), Most vision-action loop (e.g. Mid-level 2019), “Test-Time Training” (2020)
- (All of the above are test-time feedback)



Feedback

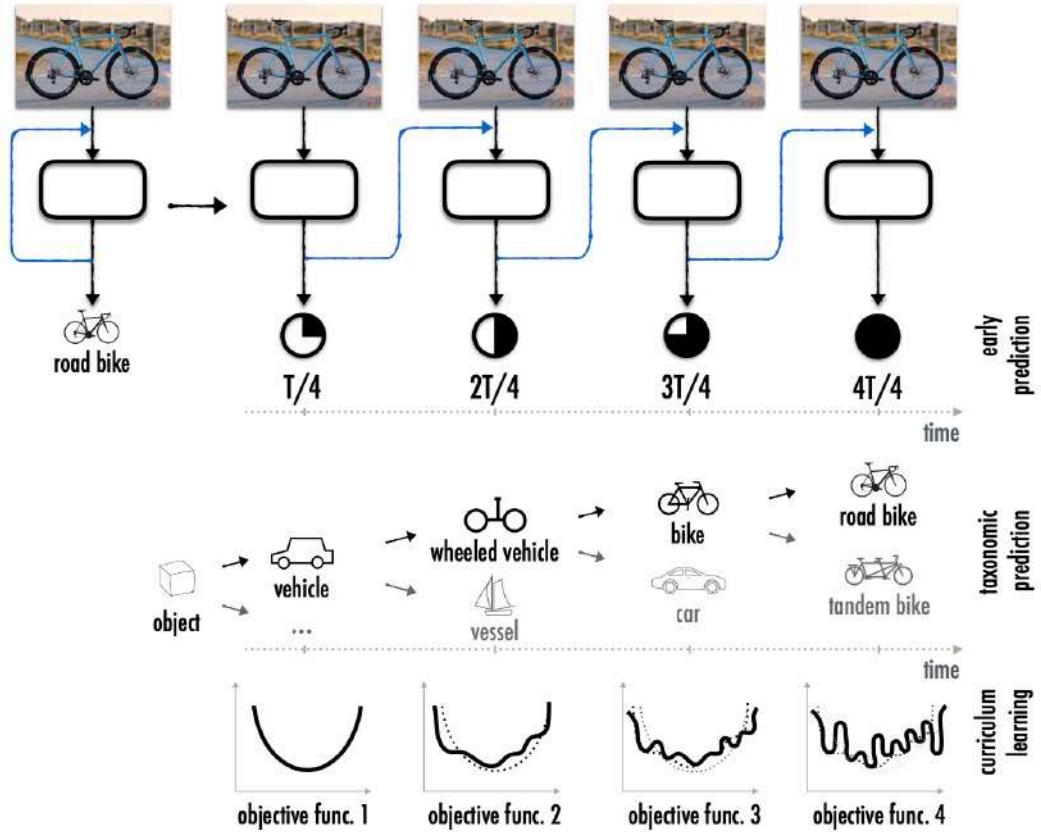


- $\epsilon_t = f(x_t)$
- $y_{t+1} = y_t + \epsilon_t$
- $x_{t+1} = I \oplus g(y_{t+1})$

$$\min_{\Theta_f, \Theta_g} \sum_{t=1}^{T} h(\epsilon_t, e(y, y_t))$$

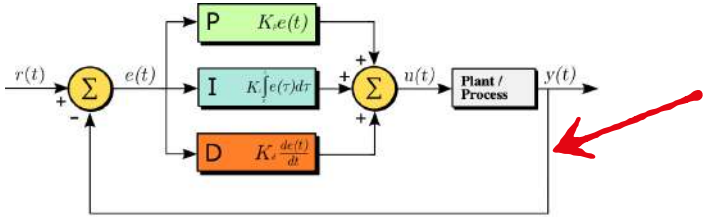
Human Pose Estimation with Iterative Error Feedback, J Carreira, P Agrawal, K Fragkiadaki, J Malik, CVPR 2016

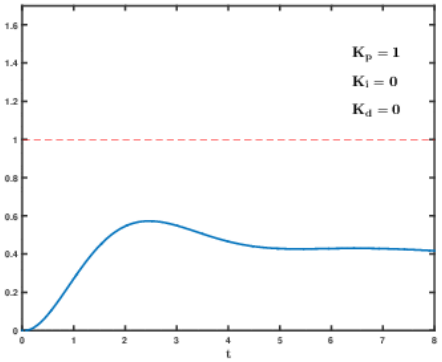
Feedback

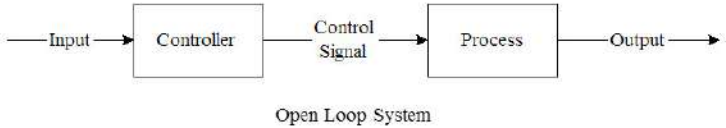

road bike
T/4
2T/4
3T/4
4T/4
early prediction
time
object
vehicle
...
wheeled vehicle
vessel
bike
car
road bike
tandem bike
taxonomic prediction
time
objective func. 1
objective func. 2
objective func. 3
objective func. 4
curriculum learning

Outer loop Feedback

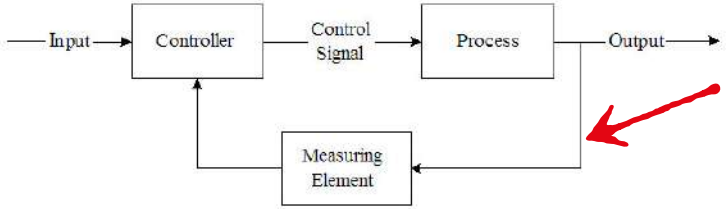
- E.g. PID controller





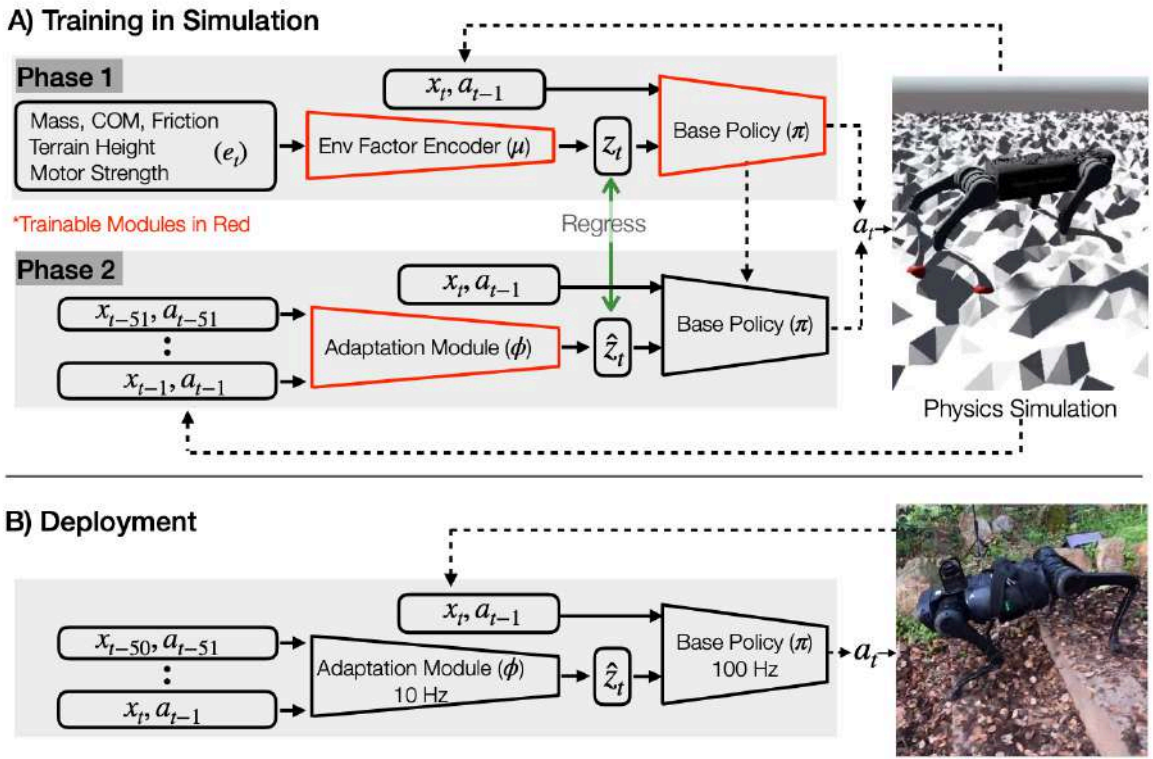


Open Loop System



Closed Loop System

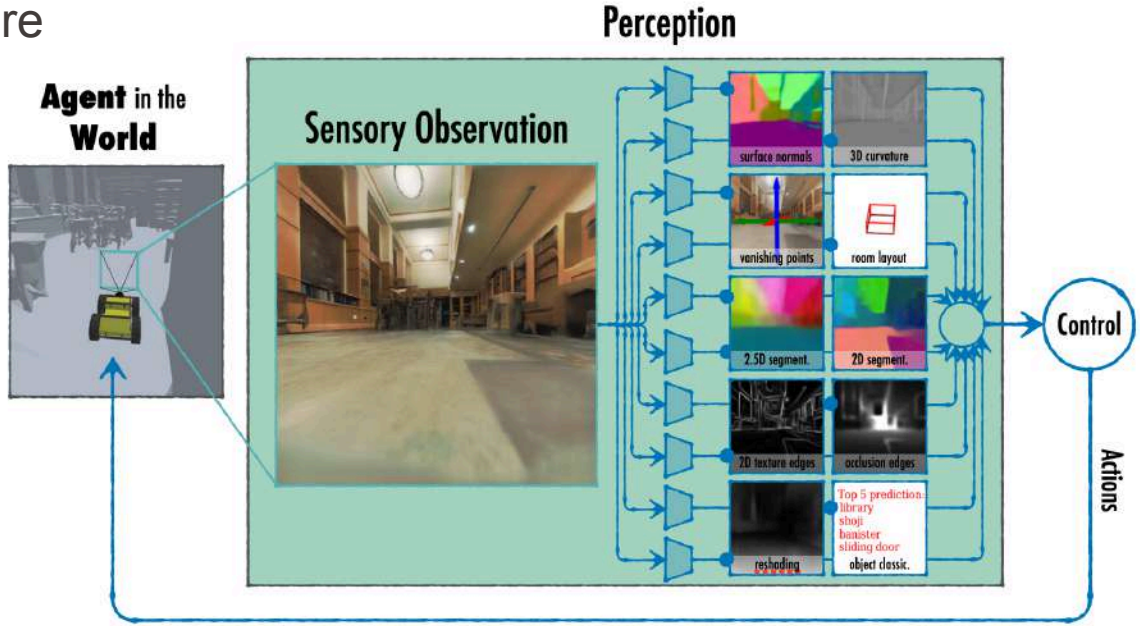
Outer loop Feedback



RMA: Rapid Motor Adaptation for Legged Robots, Kumar et al., RSS 2021

Outer loop Feedback

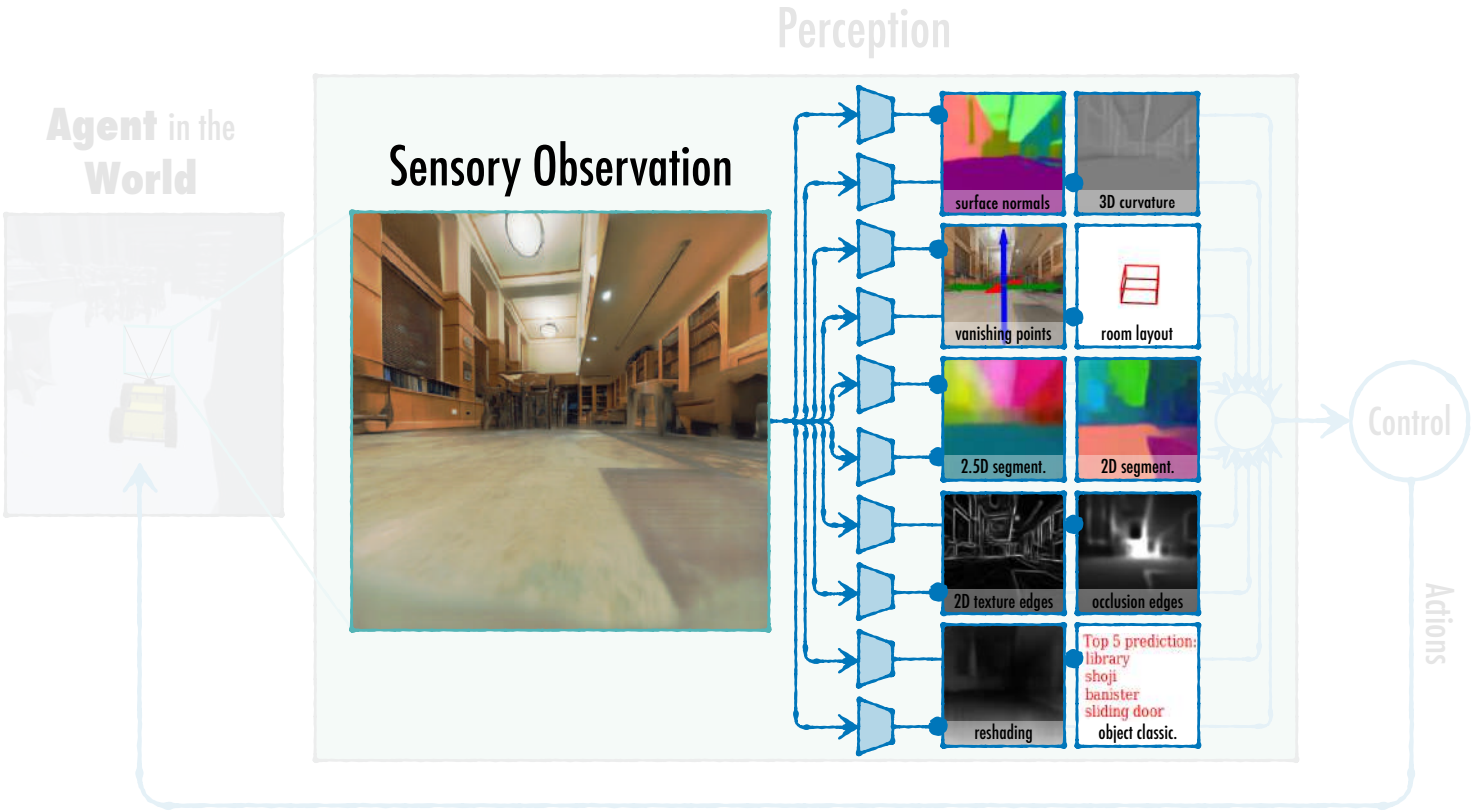
- Most vision-action systems
  - In active vision lecture



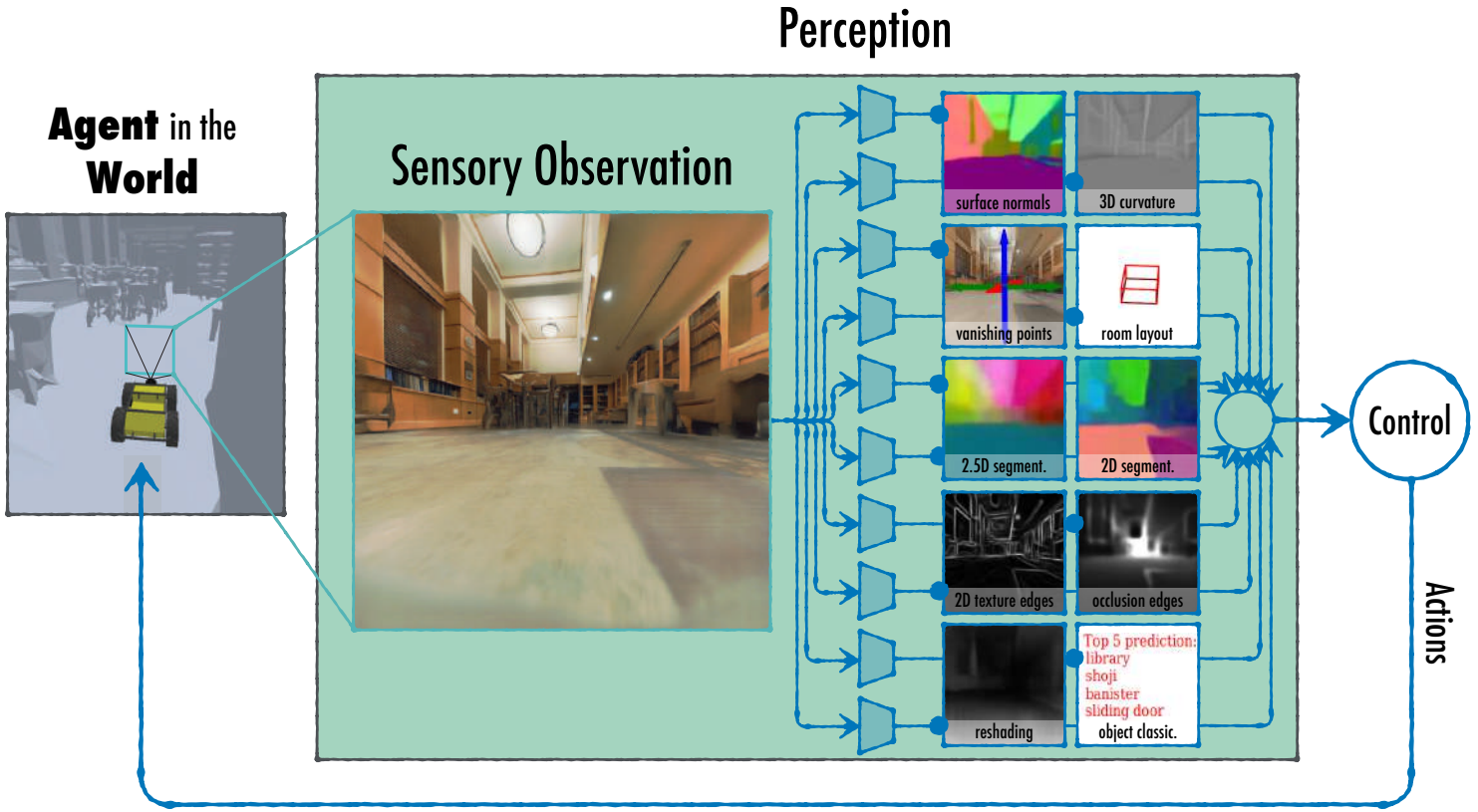
Computational Active vision

Data and Simulators

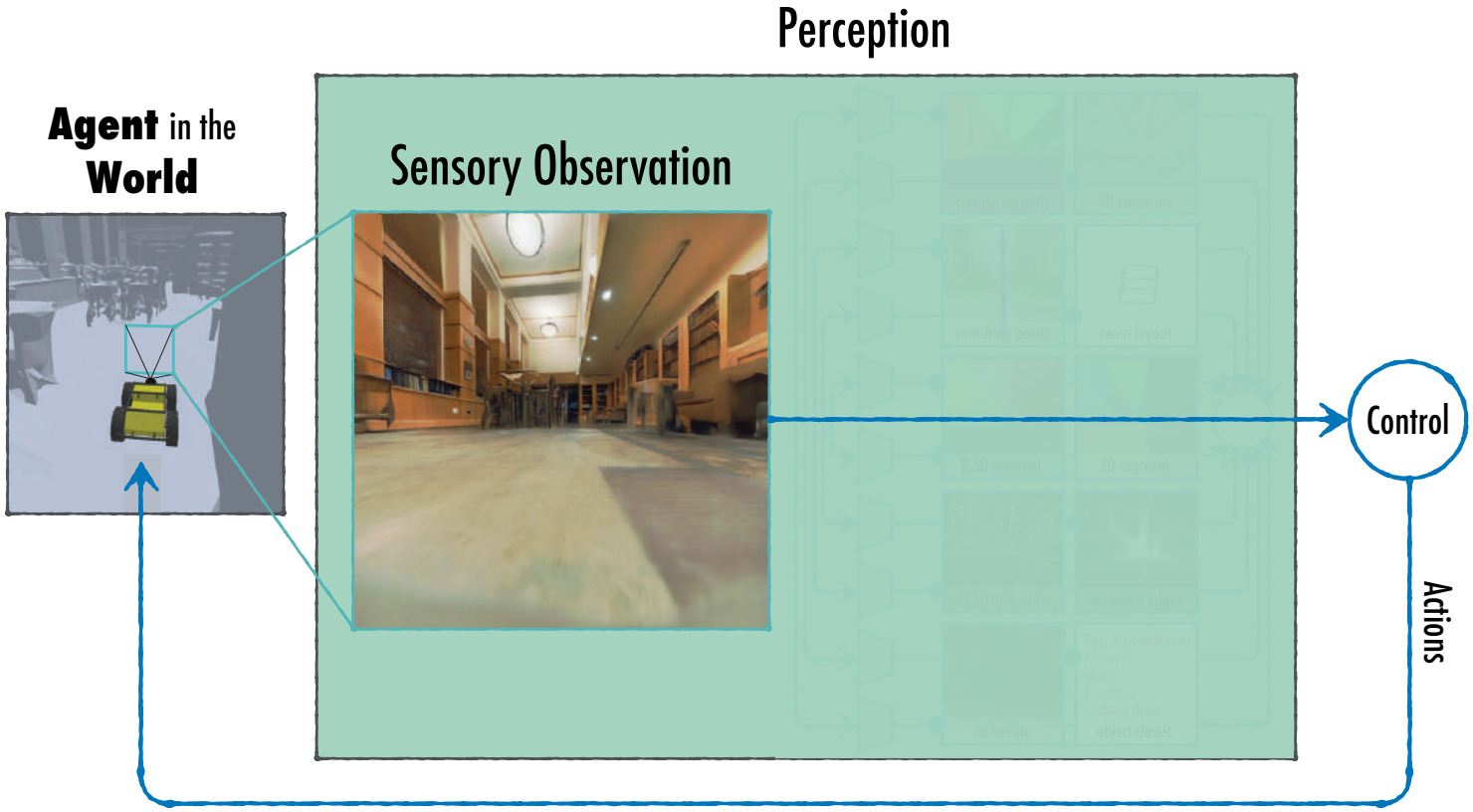
Vision In-the-loop



Vision In-the-loop



Vision In-the-loop



Dataset

Imagenet (2012)

UCF101 (2012)

Caltech101 (2004)

Berkeley Segmentation (2001)

[1] Deng, Jia, et al. "Imagenet: A large-scale hierarchical image database." CVPR 2009.
[3] Fei-Fei, Li, et al. "Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories." Computer vision and Image understanding 106, 2007
[2] Soomro, Khurram, et al. "UCF101: A dataset of 101 human actions classes from videos in the wild." arXiv 2012
[4] Martin, David, et al. "A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics." In null, 2001.
[5] Kahn, Gregory, et al. "Self-supervised Deep Reinforcement Learning with Generalized Computation Graphs for Robot Navigation", ICRA 2018

Dataset

Imagenet (2012) [1]

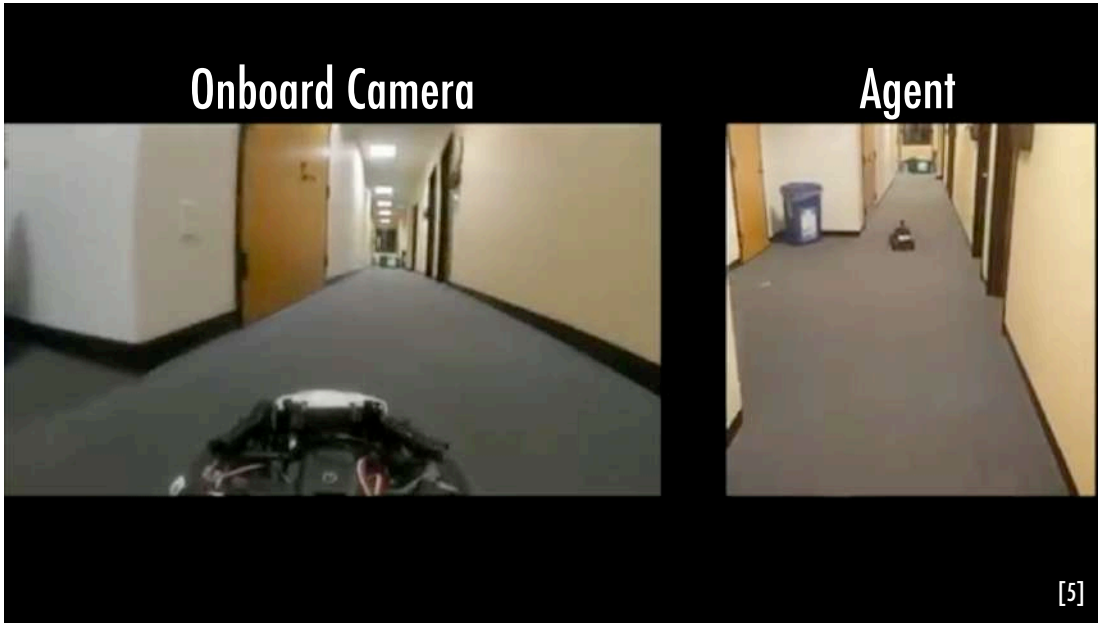
UCF101 (2012) [2]

Caltech101 (2004) [3]

Berkeley Segmentation (2001) [4]

- Passive

Perception for Active Agents



[1] Deng, Jia, et al. "Imagenet: A large-scale hierarchical image database." CVPR 2009.
[2] Soomro, Khurram, et al. "UCF101: A dataset of 101 human actions classes from videos in the wild." arXiv 2012
[3] Fei-Fei, Li, et al. "Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories." Computer vision and Image understanding 106, 2007
[4] Martin, David, et al. "A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics." In null, 2001.
[5] Kahn, Gregory, et al. "Self-supervised Deep Reinforcement Learning with Generalized Computation Graphs for Robot Navigation", ICRA 2018

Dataset

Imagenet (2012) [1]

UCF101 (2012) [2]

Caltech101 (2004) [3]

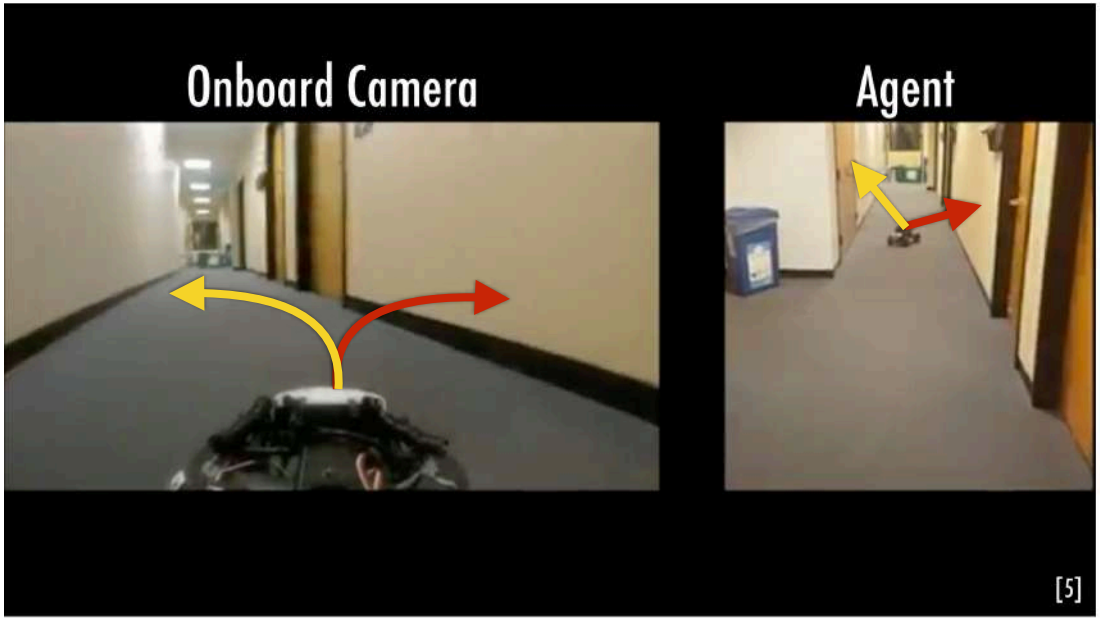
Berkeley Segmentation (2001) [4]

- Passive

[1] Deng, Jia, et al. "Imagenet: A large-scale hierarchical image database." CVPR 2009.
[3] Fei-Fei, Li, et al. "Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories." Computer vision and Image understanding 106, 2007

Perception for Active Agents

Visual observation conditioned on agent's **actions**.



[2] Soomro, Khurram, et al. "UCF101: A dataset of 101 human actions classes from videos in the wild." arXiv 2012
[4] Martin, David, et al. "A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics." In null, 2001.
[5] Kahn, Gregory, et al. "Self-supervised Deep Reinforcement Learning with Generalized Computation Graphs for Robot Navigation", ICRA 2018

EPFL

Dataset

Imagenet (2012)

UCF101 (2012)

Caltech101 (2004)

Berkeley Segmentation (2001)

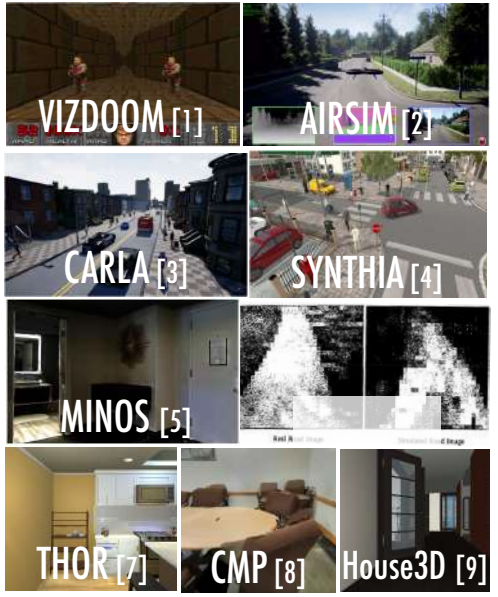
- Passive

Perception for Active Agents

Learning in physical world

- Speed bounded to real time
- Rare critical scenarios discounted
- Space bounded

Video Games & Simulators



- Generalization

[1] Kempka, Michał, et al. "Vizdoom: A doom-based ai research platform for visual reinforcement learning." CIG 2016.
[2] Shah, Shital, et al. "Airsim: High-fidelity visual and physical simulation for autonomous vehicles.", Field and service robotics 2018.
[3] Dosovitskiy, Alexey, et al "CARLA: An open urban driving simulator." arXiv 2017
[4] Ros, German, et al. "The synthia dataset: A large collection of synthetic images for semantic segmentation of urban scenes." CVPR 2016.
[5] Savva, Manolis, et al. "MINOS: Multimodal indoor simulator for navigation in complex environments." arXiv 2017.
[6] Pomerleau, Dean A. "Alvinn: An autonomous land vehicle in a neural network." Advances in neural information processing systems, 1989.
[7] Zhu, Yuke, et al. "Target-driven visual navigation in indoor scenes using deep reinforcement learning." ICRA 2017.
[8] Gupta, Saurabh, et al."Cognitive Mapping and Planning for Visual Navigation". CVPR 2017
[9] Wu, Yi, et al. "Building generalizable agents with a realistic and rich 3d environment." arXiv 2018.
[10] Pinto, Lerrel, et al. "Supersizing self-supervision: Learning to grasp from 50k tries and 700 robot hours." ICRA 2016.
[11] Gupta, Abhinav,et a;. "Robot learning in homes: Improving generalization and reducing dataset bias." In NIPS, 2018.

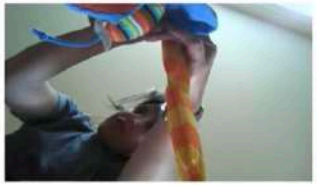
Multisensory

High resolution but naturalistic

Toyroom

Free-flowing, cluttered
In the lab

Homeview



Indiaview

At home, everyday experience

Infants: 1 month to 24 months of age

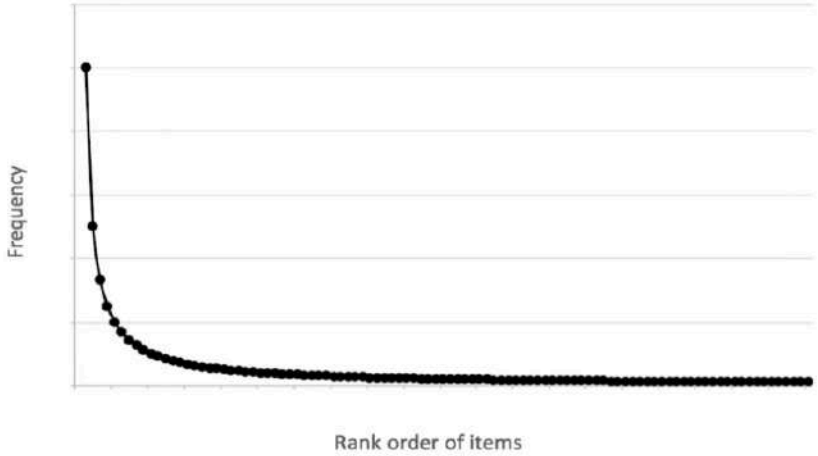
Yoshida & Smith (2008) *Infancy*
Yu et al (2009) *IEEE Transactions on Autonomous Mental Development*
Pereira, James, Jones & Smith (2010) *Journal of Vision*
Smith, Yu, Pereira, (2011) *Developmental Science*
Street, James, Jones & Smith (2011) *Child Development*
Yu & Smith (2012) *Cognition*
Yurovsky, Smith & Yu (2013) *Developmental Science*
Yu & Smith *(2013) PLoS One*
James, Swain, Jones & Smith (2014) *JCD*
Pereira, Smith & Yu (2014), *Psychological Bulletin & Review*
James et al (2014) *Developmental Science*
Jayaraman, Fausey & Smith (2015) PLoS One
Fausey , Jayaraman & Smith, (2016) *Cognition*

In aggregate 1000 hours of head –camera video, 100s millions images extracted

Yu & Smith (2016) *Current Biology*
Clerkin, Hart, Rehg, Yu & Smith (2017) *Royal Society B*
Jayaraman, Fausey & Smith (2017) *Developmental Psychology*
Suanda, Smith& Yu, (2017) *Developmental Neuropsychology*
Yu. & Smith (2017) *Child Development*
Smith, Jayaraman, Clerkin & Yu (2018) *Trends in Cognitive Science*
Jayraman & Smith (2018) *Vision Research*
Li et al (2017) ICML
Bambach,Yu,Smith & Crandall(2018) NeurIPS
Slone , Smith & Yu (2019) *Developmental Science*
McQuillian et al *(2019) Child Development*
Yuan et al (2019) *JECP*
Yuan et al (2020) *Cognition*

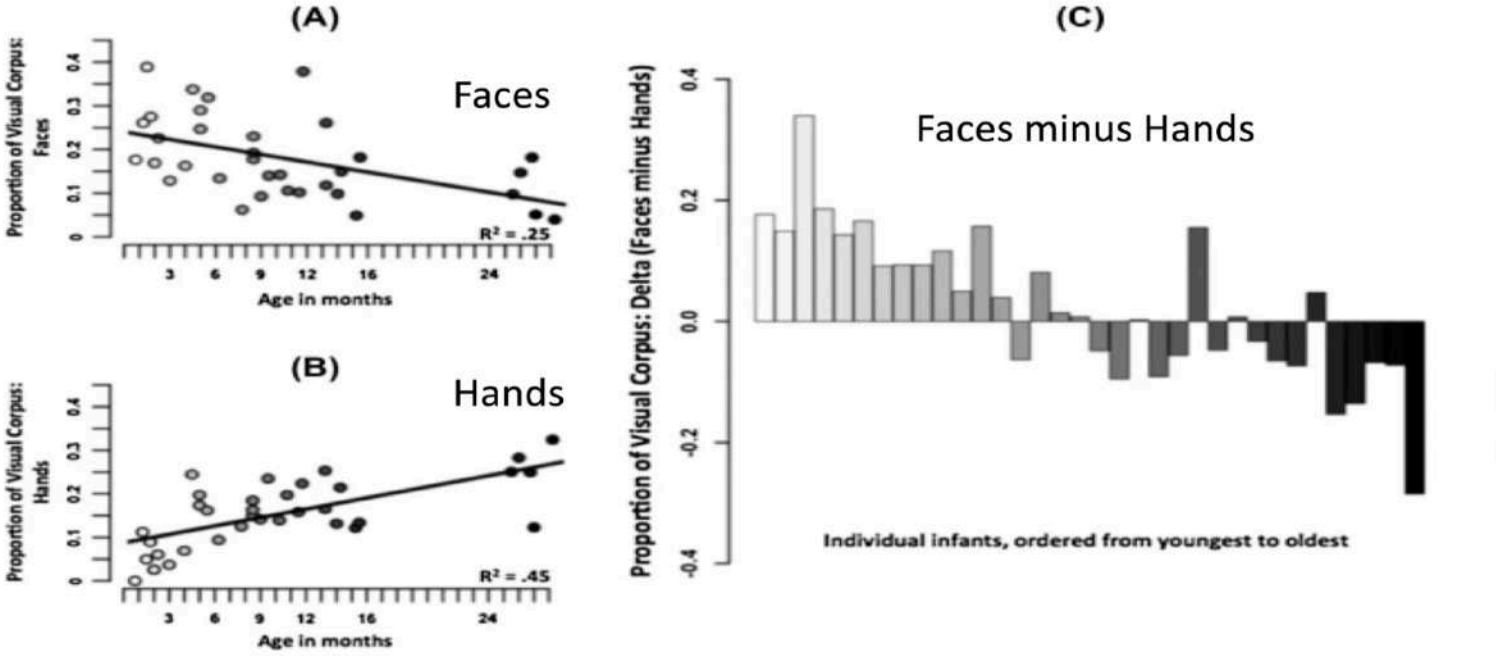
www.iub.edu/~cogdev

Not normal
Not random
Not uniform



Jayaraman S., Fausey C. & Smith LB (2015) The Faces in Infant-Perspective Scenes Change over the First Year of Life. **PLoS ONE**, 10(5).
Clerkin, E. M., Hart, E., Rehg, J. M., Yu, C. & Smith, L. B. (2017) Real-World Visual Statistics and Infants' First-learned Object Names. Philosophical **Transactions of the Royal Society B**, 372.
Slone, L., Smith, L.B., Yu, C (2019). Self-generated variability in object images predicts vocabulary growth. **Developmental Science**

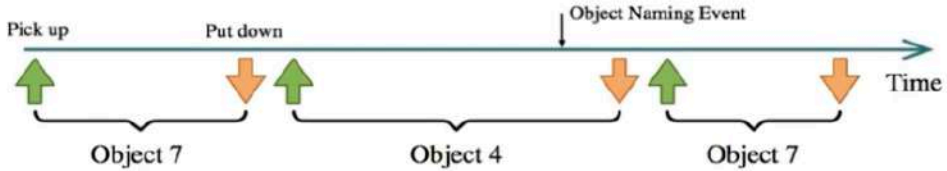

Frequency
Rank order of items



Fausey ,Jayaraman & Smith, (2016) *Cognition*

Overlapping “tasks” without clear definition

what is the infant’s task when handling an object.



Invited Talk

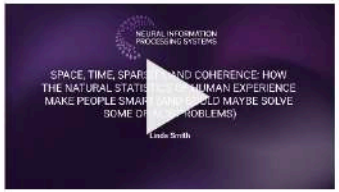
Coherence statistics, self-generated experience and why young humans are much smarter than current AI.

Linda Smith

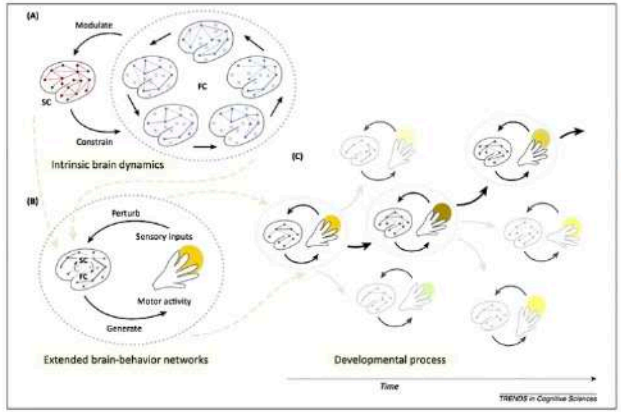
Hall E (level 1)

[Abstract]

Tue 12 Dec 12:15 p.m. PST — 1:05 p.m. PST



The only training data that matters is the data received by the **child's** sensors



It's a physical world

Therefore, the received input is constrained by **time and space**

The **spatial** relations of sensors to the world determine-- **moment to moment** – the received data

This **brain-behavior-input loop** is essential to the adaptation and learning in mammalian brains

Byrge, L., Sporns, O. & Smith, L. B. (2014) Developmental process emerges from extended brain-body-behavior networks. *Trends in Cognitive Sciences*

NeurIPS 2024

Gibson Environment: Virtualizing Real Spaces

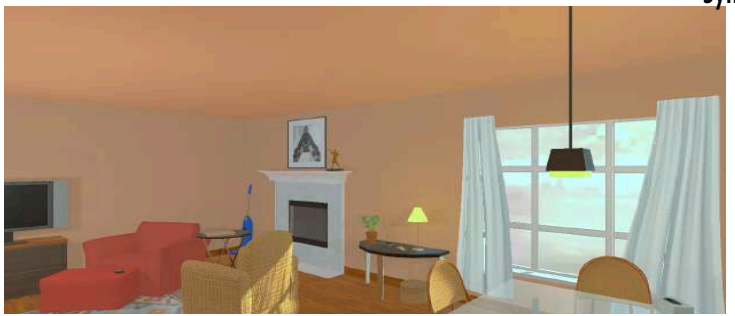
[1] Matterport3D,
[2] Kinect,
[3] Google Tango,
[4] FARO, [6] NavVis,
[5] Newcombe, et al. "KinectFusion: Real-time dense surface mapping and tracking", 2011.
[6] Dai, A., et al. Bundlefusion: Real-time globally consistent 3d reconstruction using on-the-fly surface reintegration. ACM ToG 2017
[7] Durrant-Whyte, H.et al. "Simultaneous localization and mapping: part I". Robotics & Automation Magazine. 2006

"Gibson Env: Real-World Perception for Embodied Agents". Xia*, Zamir*, He*, Sax, Malik, Savarese. CVPR 2018. [NVIDIA Pioneering Research Award]

Gibson Environment: Virtualizing Real Spaces

Virtualized

Synthetic



[1] Kempka, Michał, et al. "Vizdoom: A doom-based ai research platform for visual reinforcement learning." CIG 2016.
[2] Shah, Shital, et al. "Airsim: High-fidelity visual and physical simulation for autonomous vehicles.", Field and service robotics 2018.
[3] Ros, German, et al. "The synthia dataset: A large collection of synthetic images for semantic segmentation of urban scenes." CVPR 2016.
[4] Dosovitskiy, Alexey, et al "CARLA: An open urban driving simulator." arXiv 2017

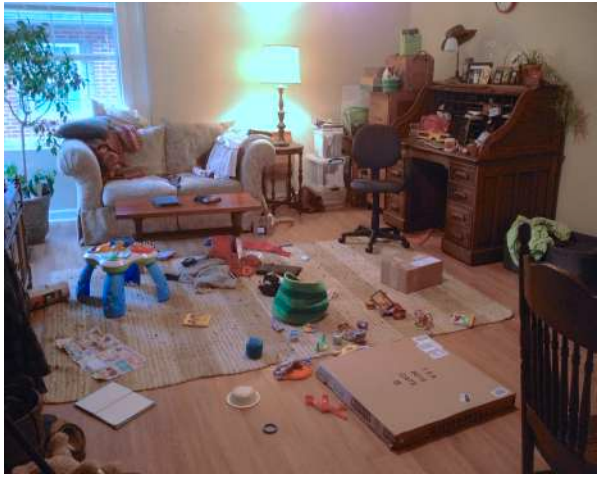
Simulation

Real-World

Generalization from Simulation to real-world

Issue I:
photorealism

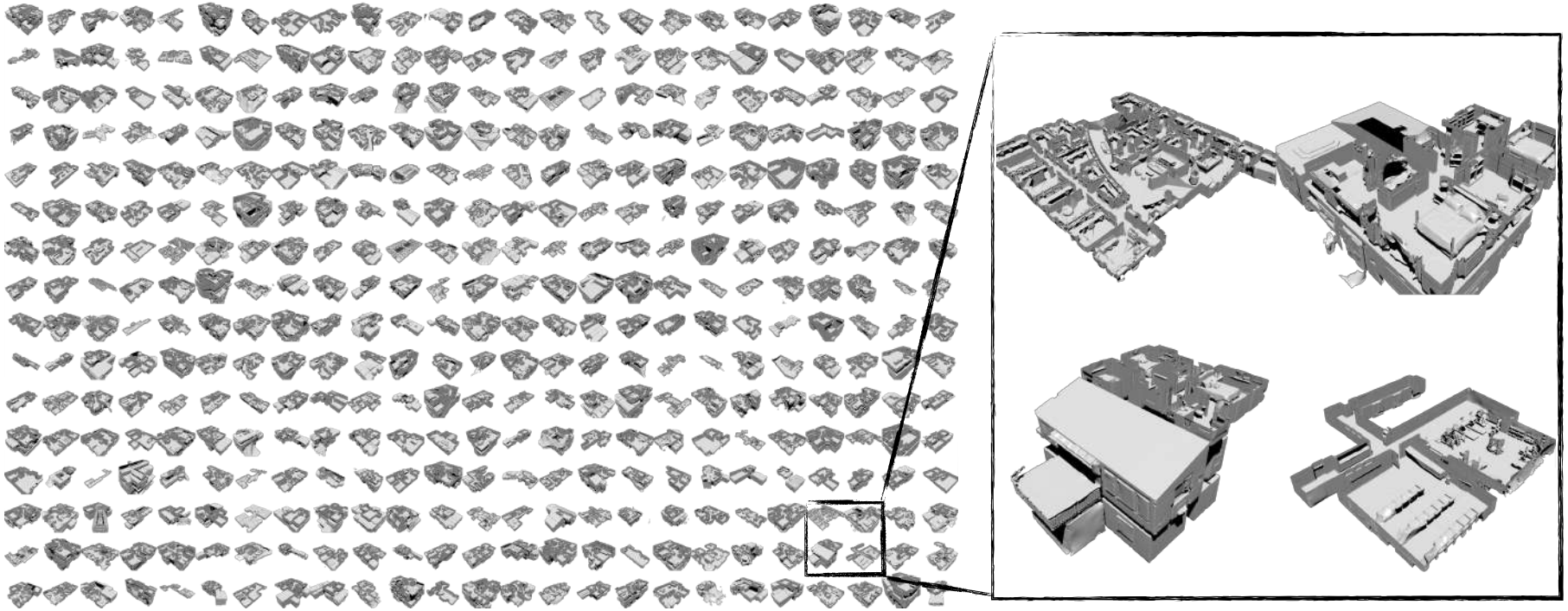
Gibson Environment



Issue II:
semantic distribution mismatch

Gibson Environment

Large Real Space

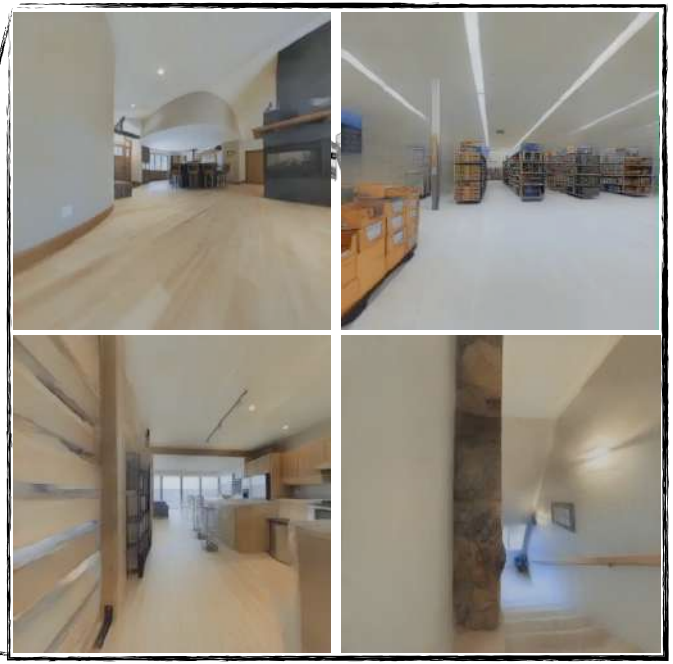


572 full buildings. Real spaces, scanned and reconstructed in 3D.

Browse data at: http://gibsonenv.stanford.edu/database/

Gibson Environment

Large Real Space

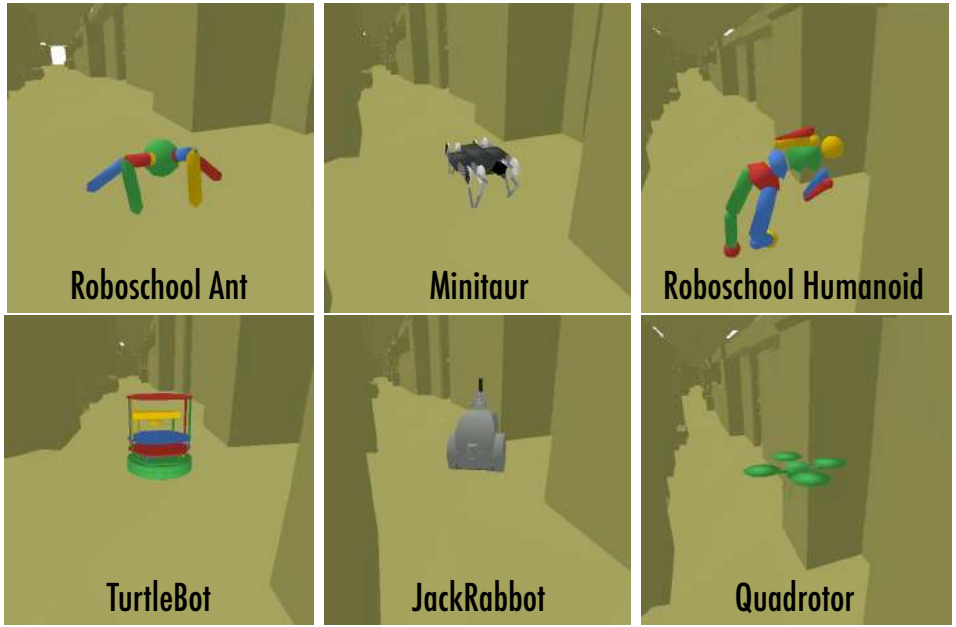


572 full buildings. Real spaces, scanned and reconstructed in 3D.

Browse data at: http://gibsonenv.stanford.edu/database/

Gibson Environment

Active Agent



Arbitrary agents can be improved using their URDF.

Gibson Environment

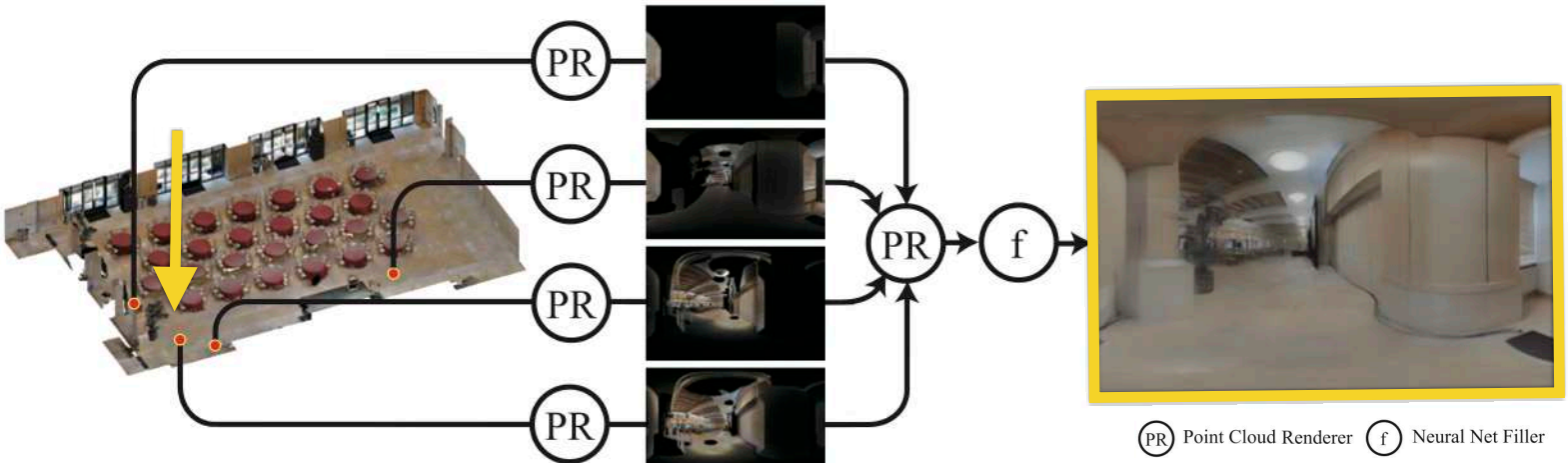
Subject to physics

Integrated with physics engine, PyBullet3D. [Coumans2016]

Gibson Environment

RGB Frame Stream

View synthesis engine



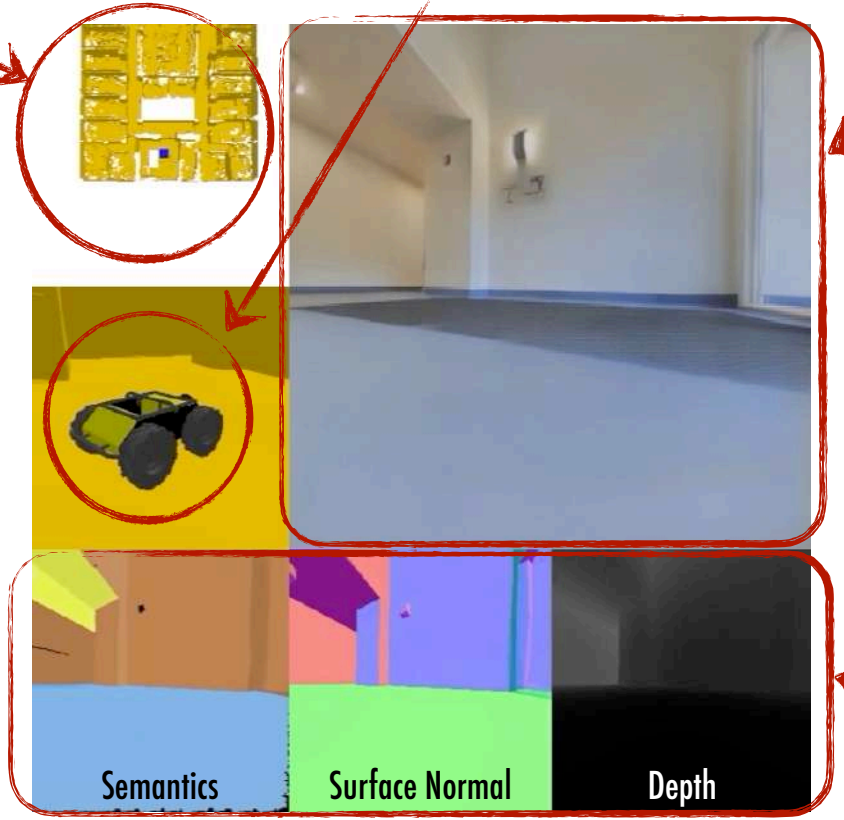
Given sparse RGB-D images, renders the scene from arbitrary viewpoints.

Gibson Environment

Large Real Space

Active Agent

RGB Frame Stream



Additional Modalities

Gibson Environment

Large Real Space

Active Agent

RGB Frame Stream

Additional Modalities

Gibson Environment

Large Real Space

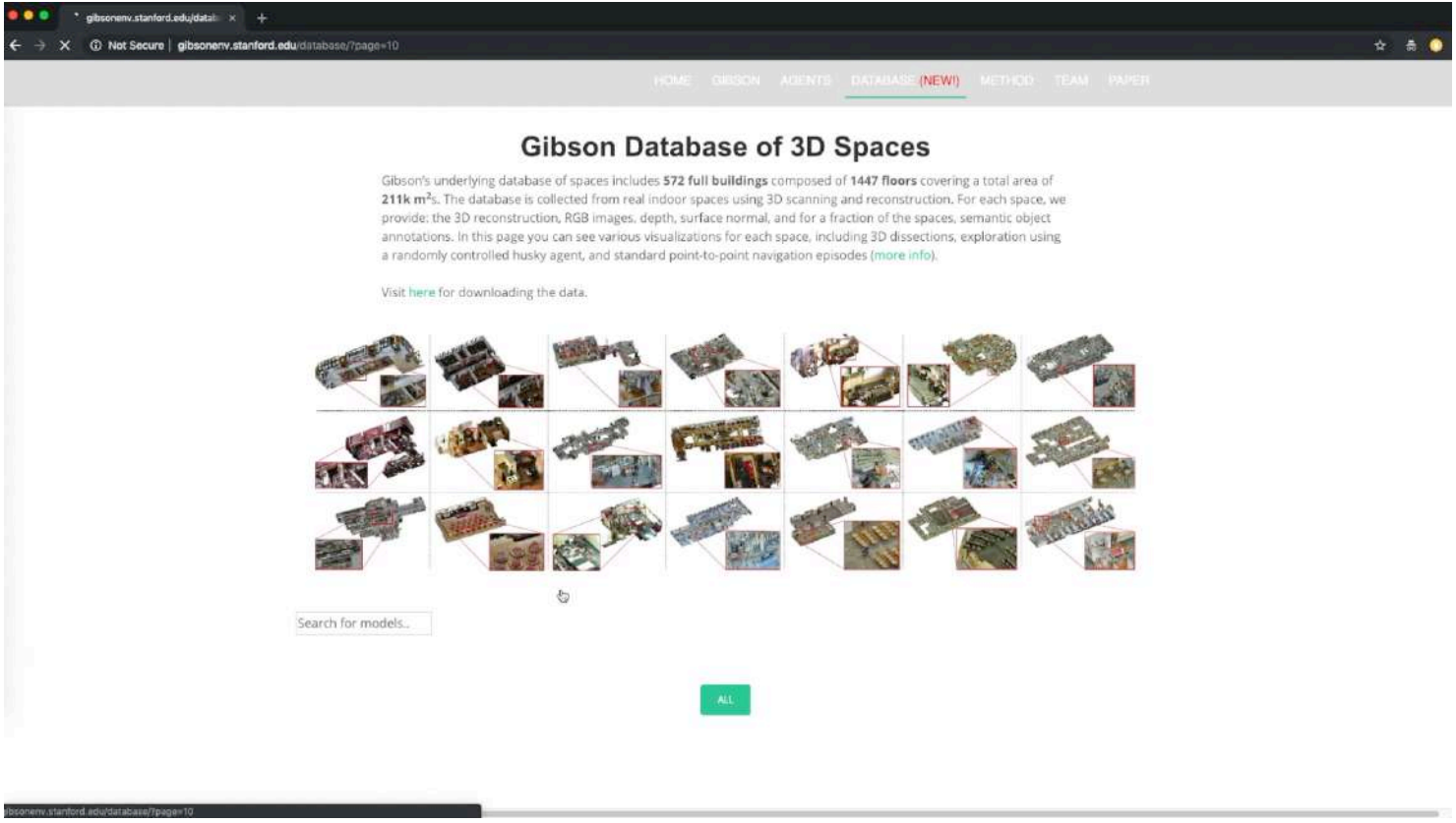
Active Agent

RGB Frame Stream

Additional Modalities

Explore the Gibson buildings

http://gibson.vision/database/

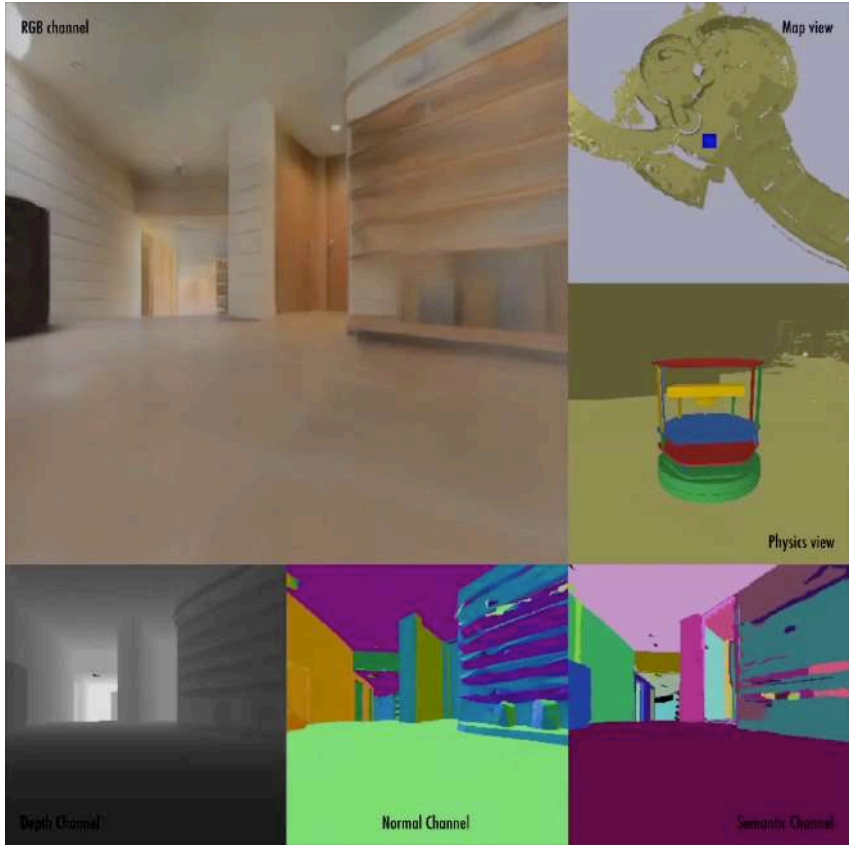


Gibson Environment at a glance

Large Real Space

Active Agent

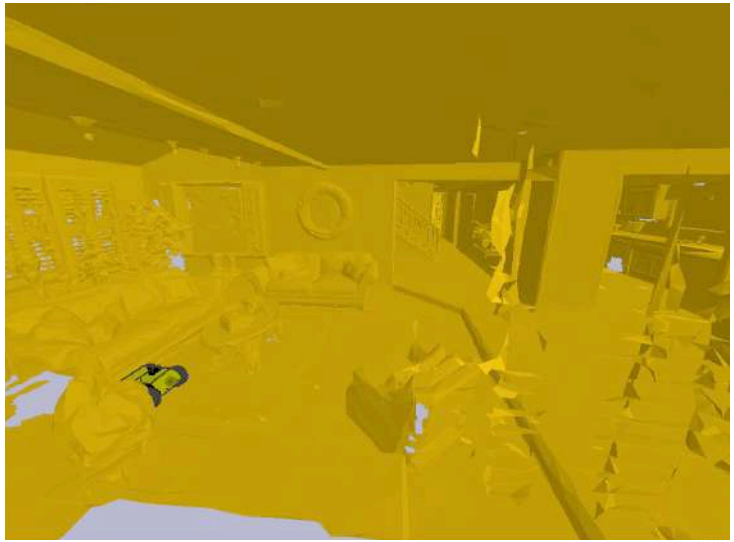
RGB Frame Stream



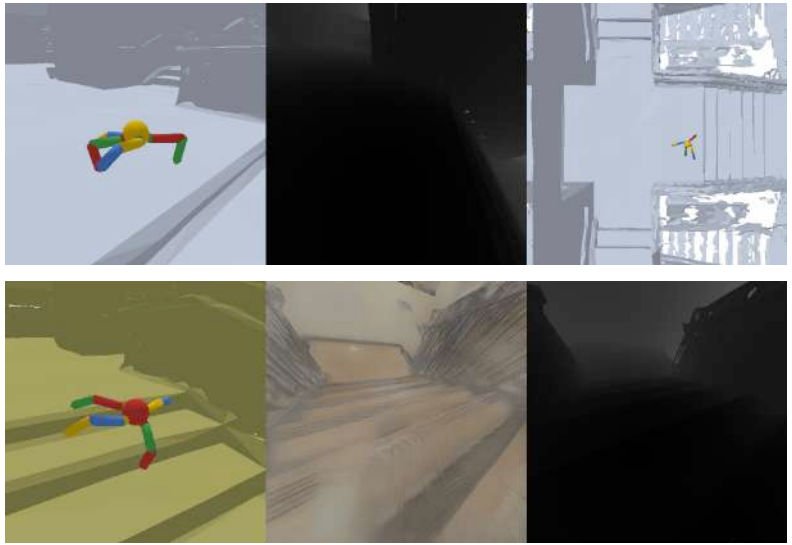
Subject to physics

Additional Modalities

Sample perceptual agents trained in Gibson (using Reinforcement Learning)



Local planning ("go to the target")



Stair climb

Community follow-ups using Gibson Environment

Generalization through Simulation:
Integrating Simulated and Real Data into Deep Reinforcement Learning for Vision-Based Autonomous Flight

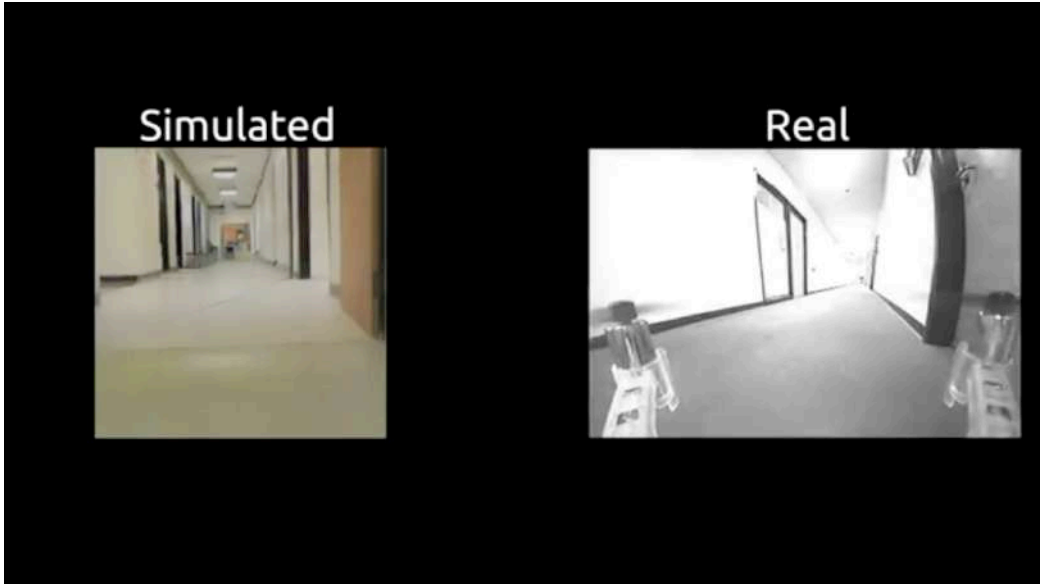
Katie Kang*, Suneel Belkhale*, Gregory Kahn*, Pieter Abbeel, Sergey Levine
Berkeley AI Research (BAIR), University of California, Berkeley

45

Community follow-ups using Gibson Environment

Generalization through Simulation:
Integrating Simulated and Real Data into Deep Reinforcement Learning for Vision-Based Autonomous Flight

Katie Kang*, Suneel Belkhale*, Gregory Kahn*, Pieter Abbeel, Sergey Levine
Berkeley AI Research (BAIR), University of California, Berkeley

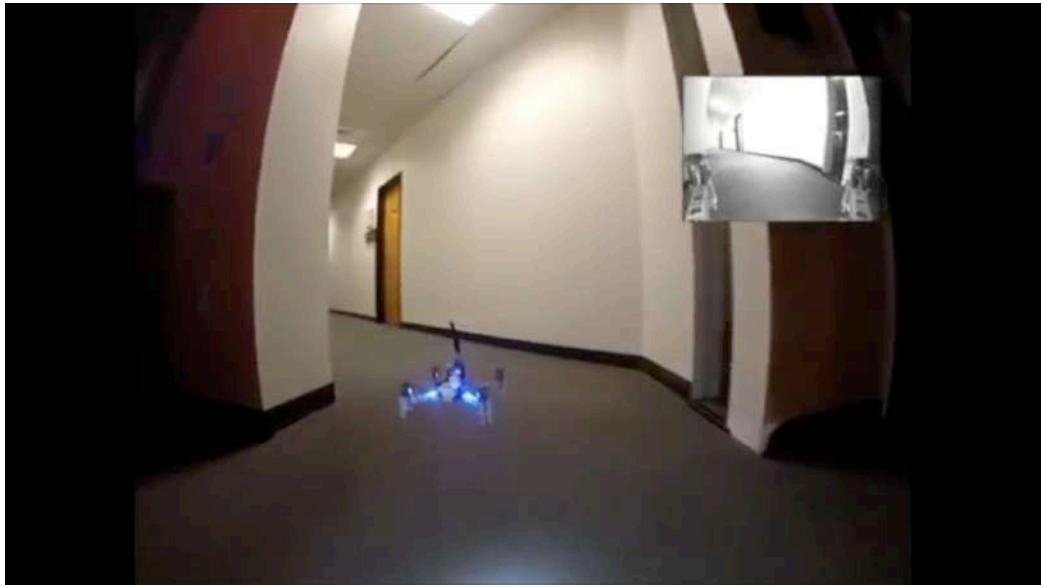


46

Community follow-ups using Gibson Environment

Generalization through Simulation:
Integrating Simulated and Real Data into Deep Reinforcement Learning for Vision-Based Autonomous Flight

Katie Kang*, Suneel Belkhale*, Gregory Kahn*, Pieter Abbeel, Sergey Levine
Berkeley AI Research (BAIR), University of California, Berkeley



47

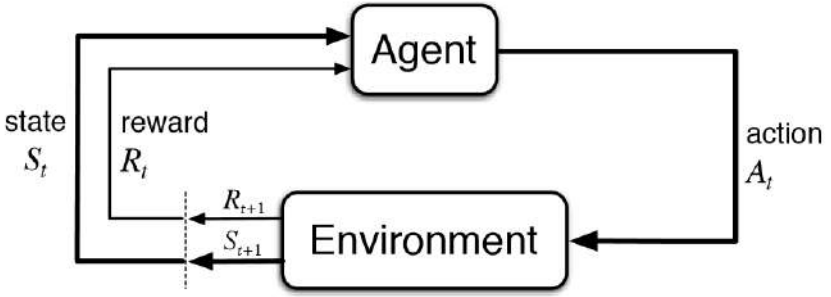
Community follow-ups using Gibson Environment

Generalization through Simulation:
Integrating Simulated and Real Data into Deep Reinforcement Learning for Vision-Based Autonomous Flight

Katie Kang*, Suneel Belkhale*, Gregory Kahn*, Pieter Abbeel, Sergey Levine
Berkeley AI Research (BAIR), University of California, Berkeley

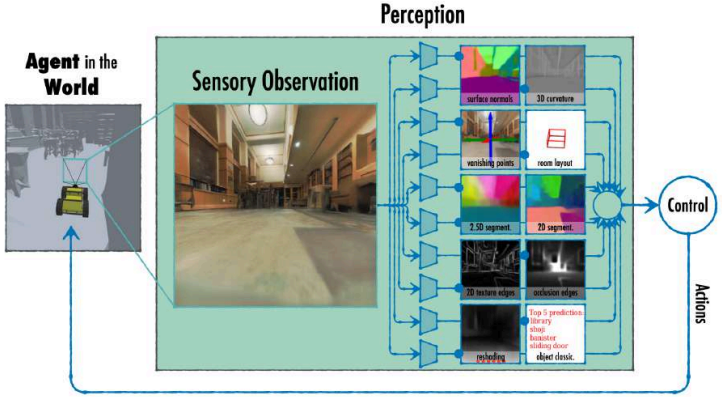
48

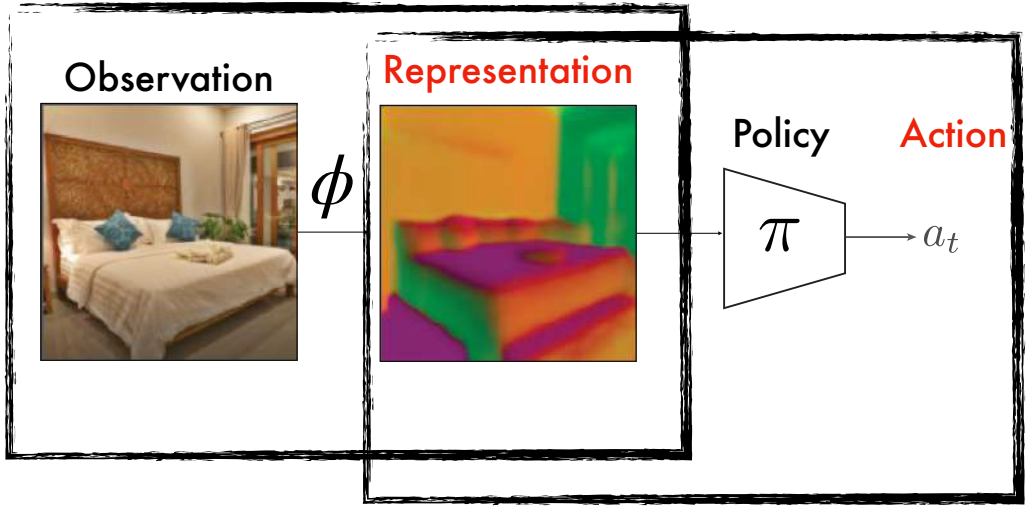
Perceptual Robotics



Zamir

1: How to **represent** the sensory information?



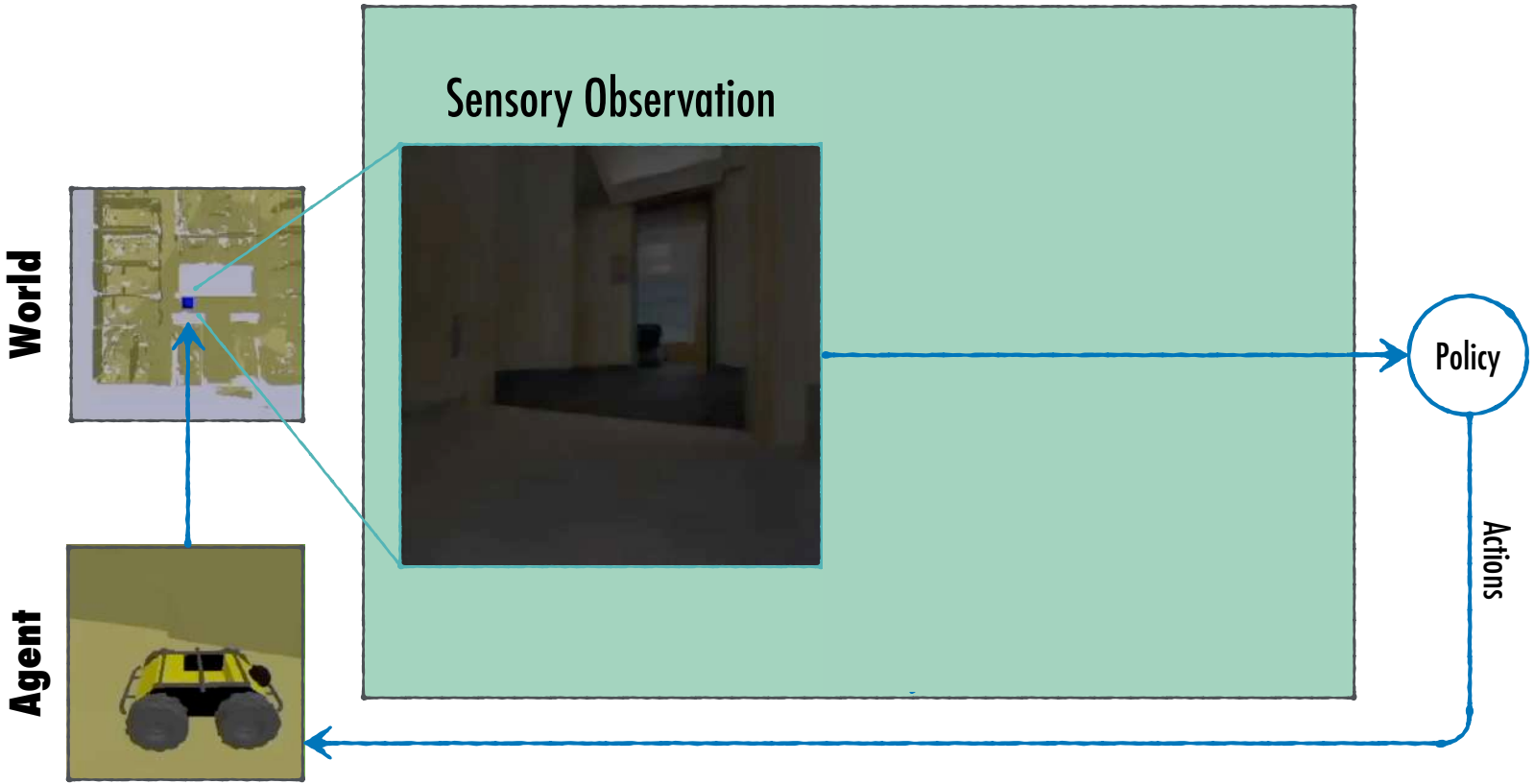


2: How to infer **actions** of the representation?

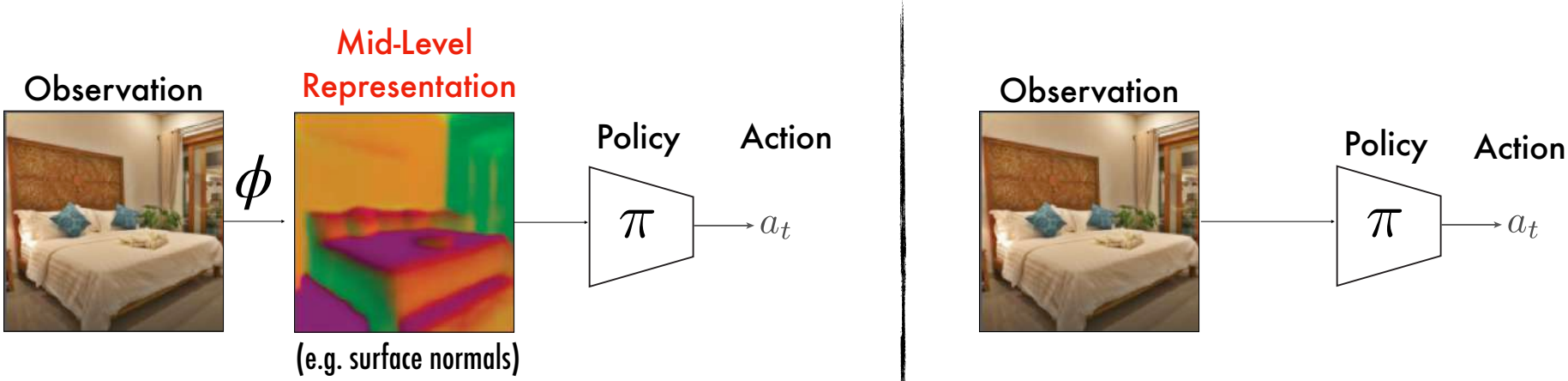
Mid-Level Vision for Robotics

Robust Policies via Mid-Level Visual Representations: An Experimental Study in Manipulation and Navigation. **Conference on Robot Learning (CoRL) '20**.
Learning to Navigate Using Mid-Level Visual Priors. **Conference on Robot Learning (CoRL) '19**
Mid-Level Visual Representations Improve Generalization and Sample Efficiency for Learning Visuomotor Policies. **(BayLearn '19)**

Vision in the action loop



Setup



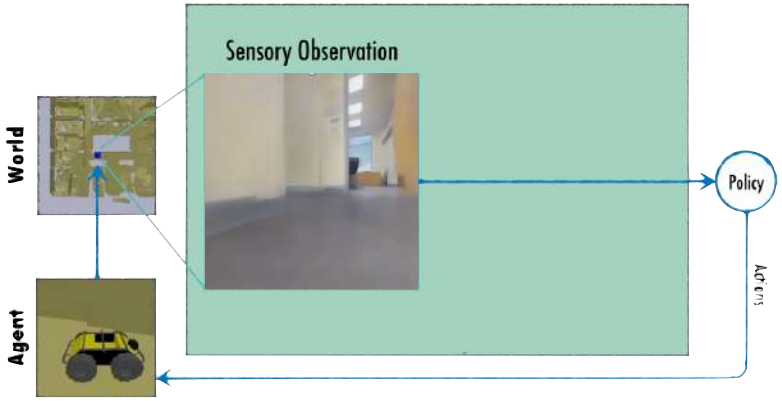
Learning with Perceptual Priors Vs "Tabula Rasa" (scratch) Learning

Tested hypothesis 1: Does mid-level vision **accelerate learning**?

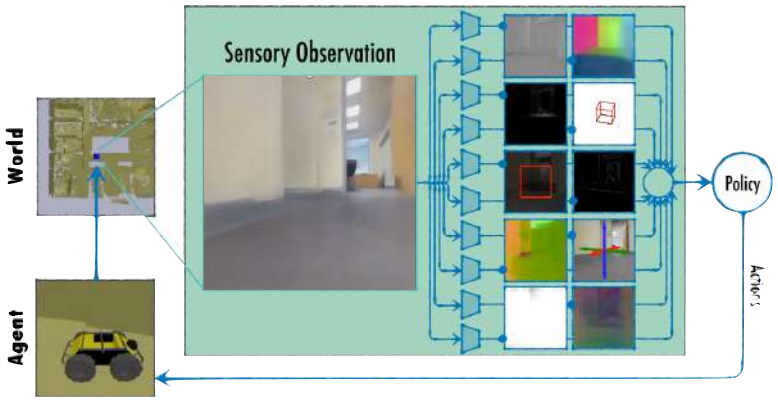
Tested hypothesis 2: Can mid-level features **generalize** better to unseen spaces?

- "Mid-Level Visual Representations Improve Generalization and Sample Complexity for Learning Visuomotor Policies". Sax, Emi, Zamir, Guibas, Savarese, Malik. Arxiv 2018. CoRL 2019.
- "Robust Policies via Mid-Level Visual Representations: An Experimental Study in Manipulation and Navigation". Chen, Sax, Pinto, Lewis, Armeni, Savarese, Zamir, Malik. CoRL 2020

Setup



"Tabula Rasa" (scratch) Learning



Learning with Perceptual Priors

Tested hypothesis 1: Does mid-level vision **accelerate learning**?

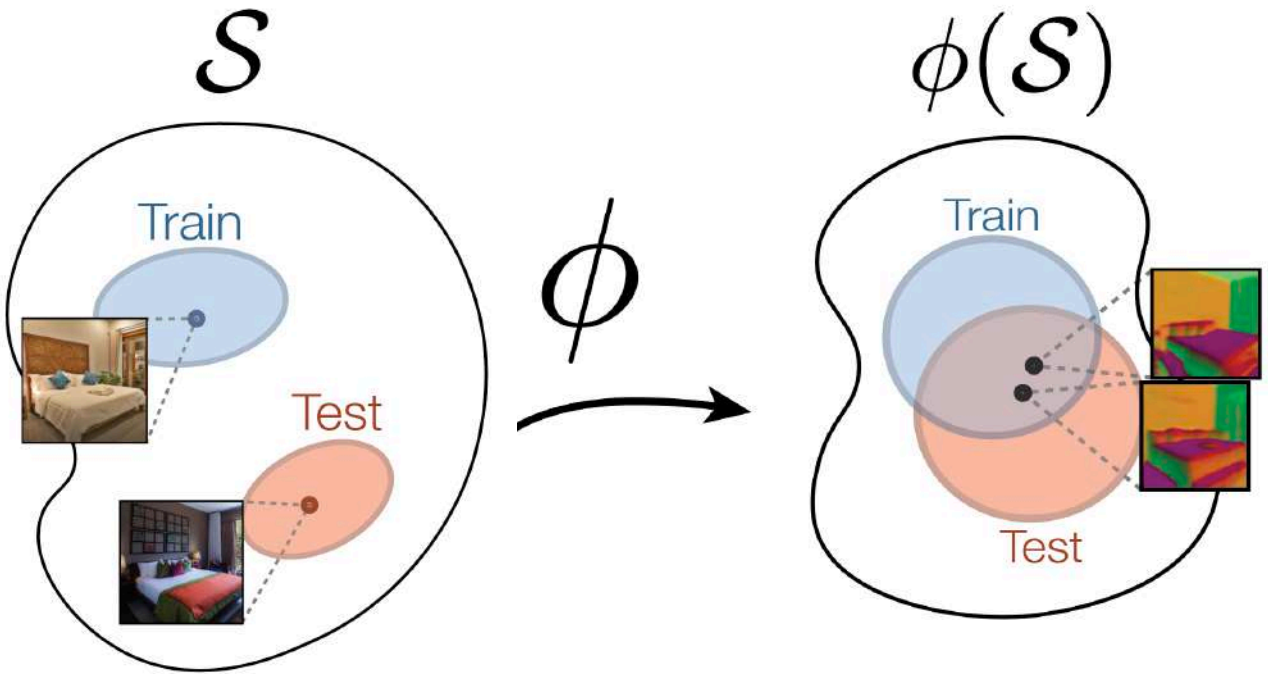
Tested hypothesis 2: Can mid-level features **generalize** better to unseen spaces?

- "Mid-Level Visual Representations Improve Generalization and Sample Complexity for Learning Visuomotor Policies". Sax, Emi, Zamir, Guibas, Savarese, Malik. Arxiv 2018. CoRL 2019.
- "Robust Policies via Mid-Level Visual Representations: An Experimental Study in Manipulation and Navigation". Chen, Sax, Pinto, Lewis, Armeni, Savarese, Zamir, Malik. CoRL 2020

Why Perceptual Priors Could Help?

Raw Sensory Data (pixel) Space

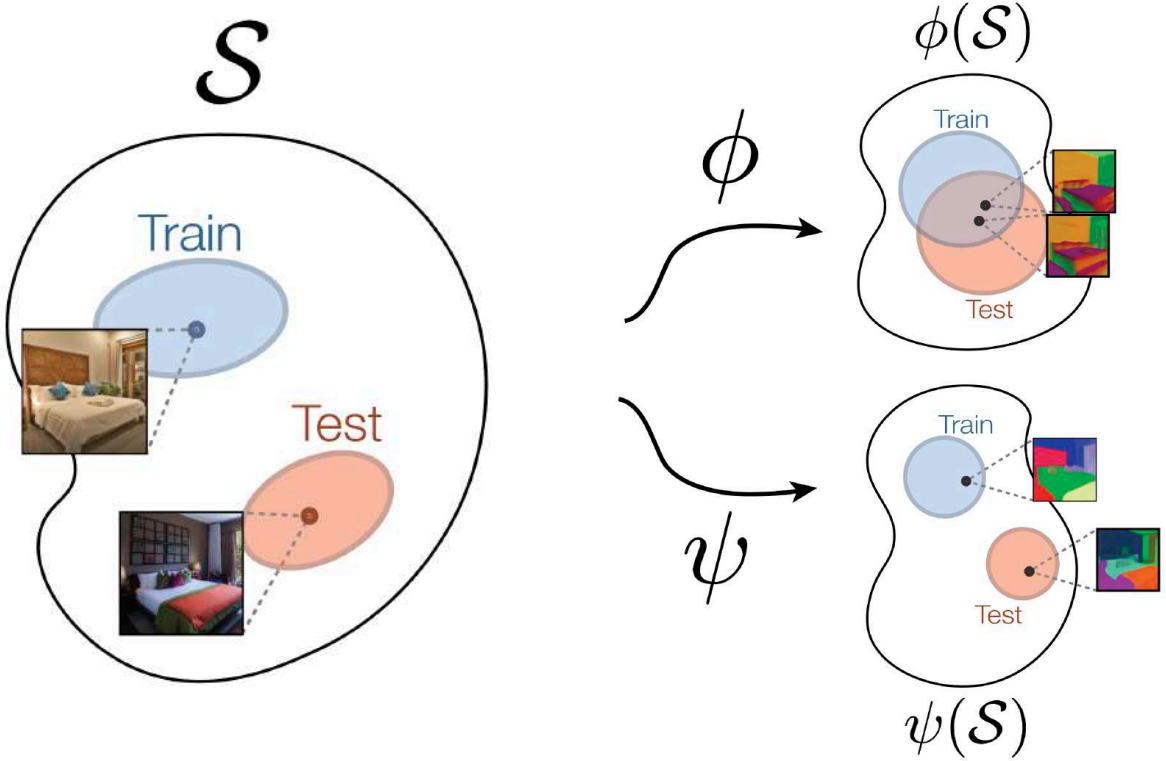
Visual Feature Space



Why Perceptual Priors Could Help?

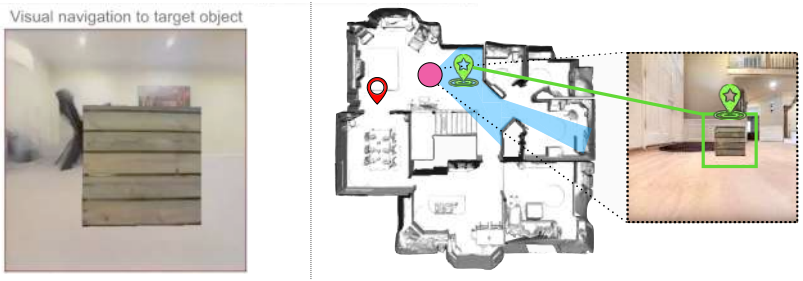
Raw Sensory Data (pixel) Space

Visual Feature Space



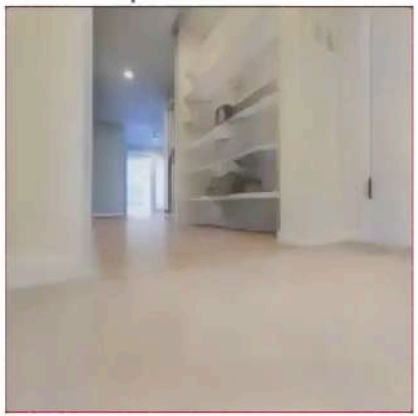
Downstream Active Tasks

Visual Navigation



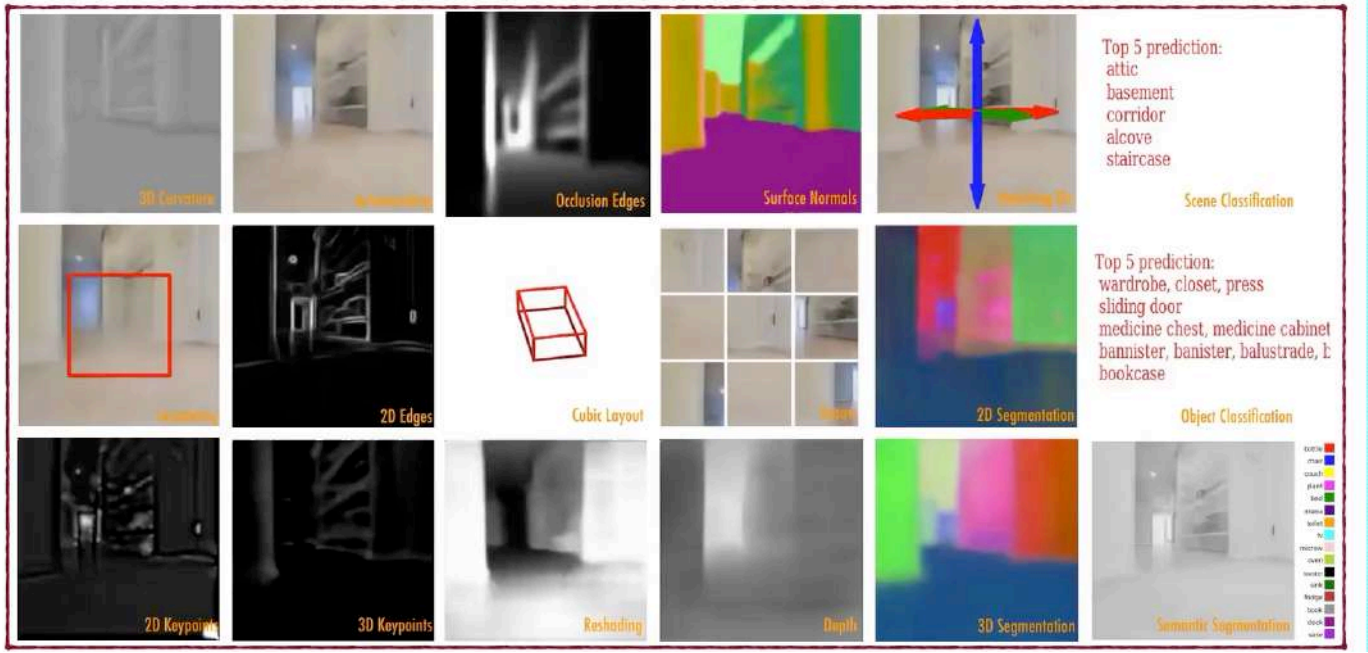
Mid-Level Features

Input Frames



Husky robot in Gibson navigating to target object

Mid-Level Vision



Learning with vs without perceptual priors

Visual navigation to target object

With mid-level vision

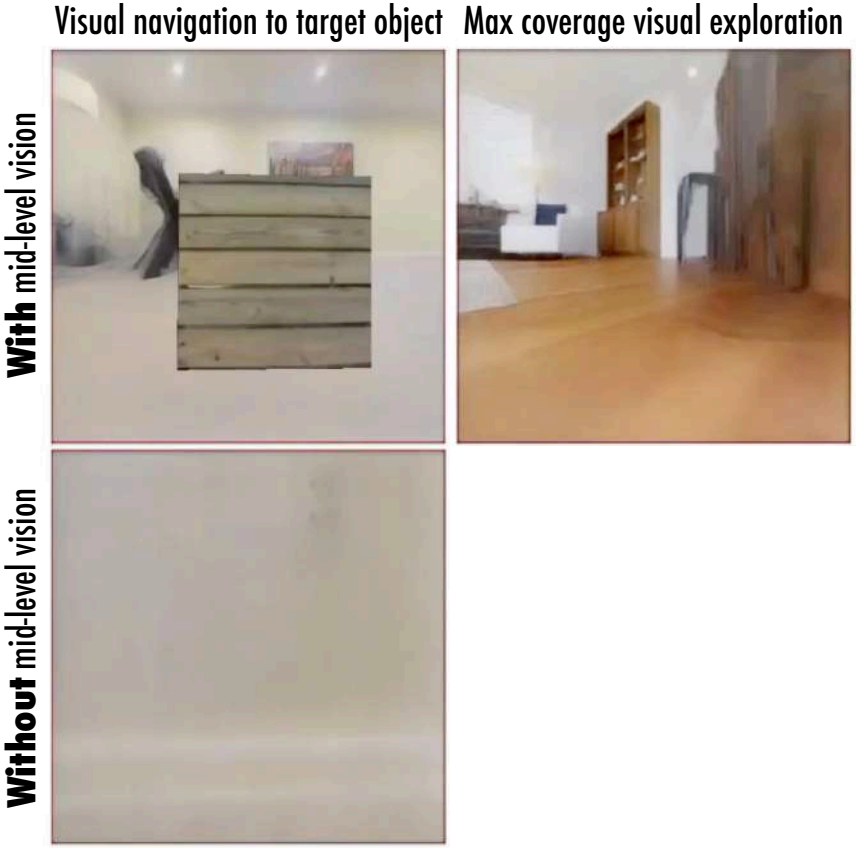
Learning with vs without perceptual priors

Visual navigation to target object

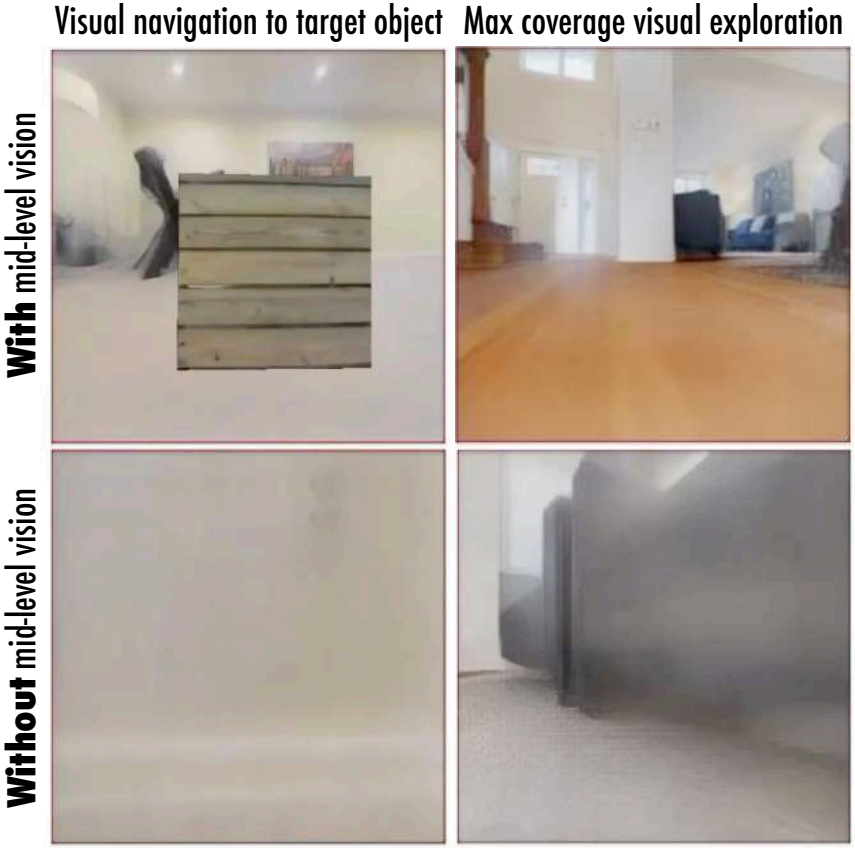
With mid-level vision

Without mid-level vision

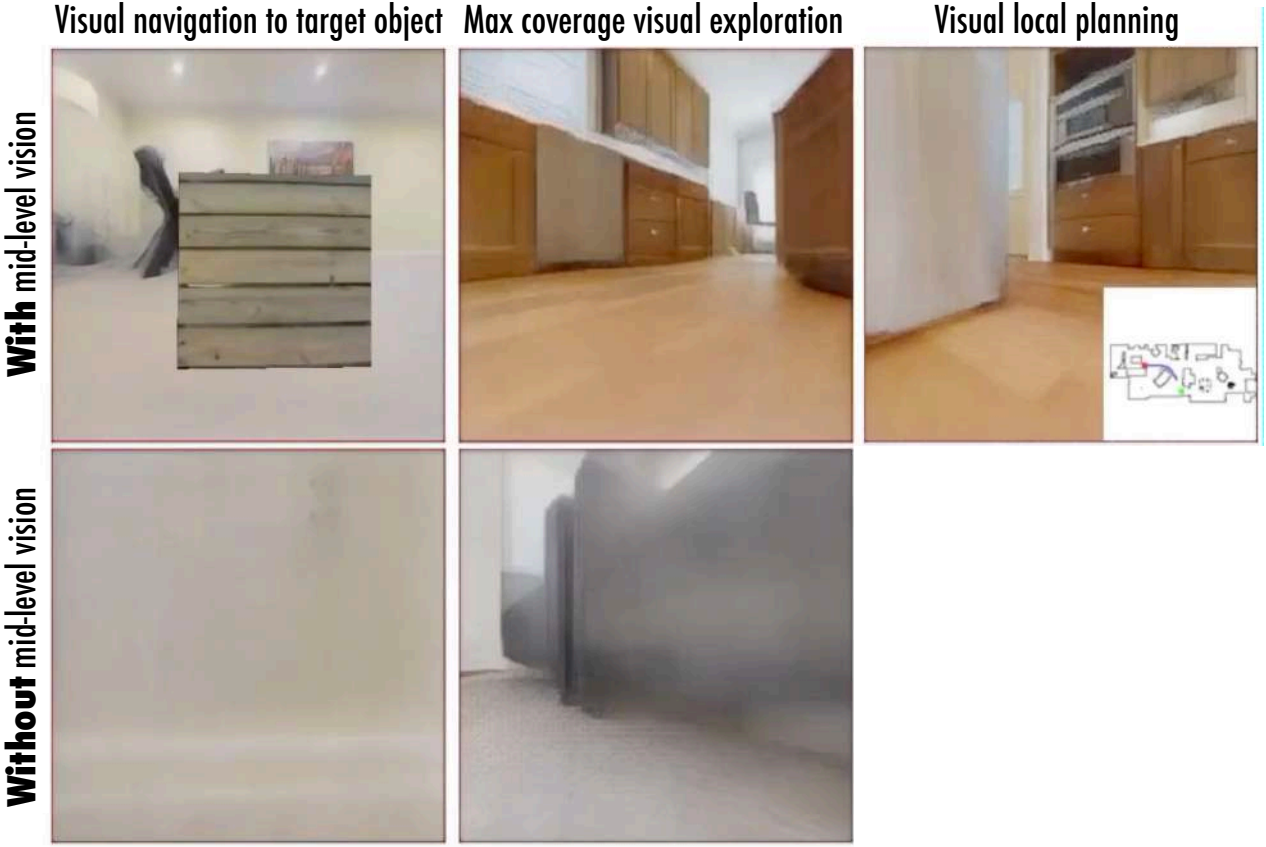
Learning with vs without perceptual priors



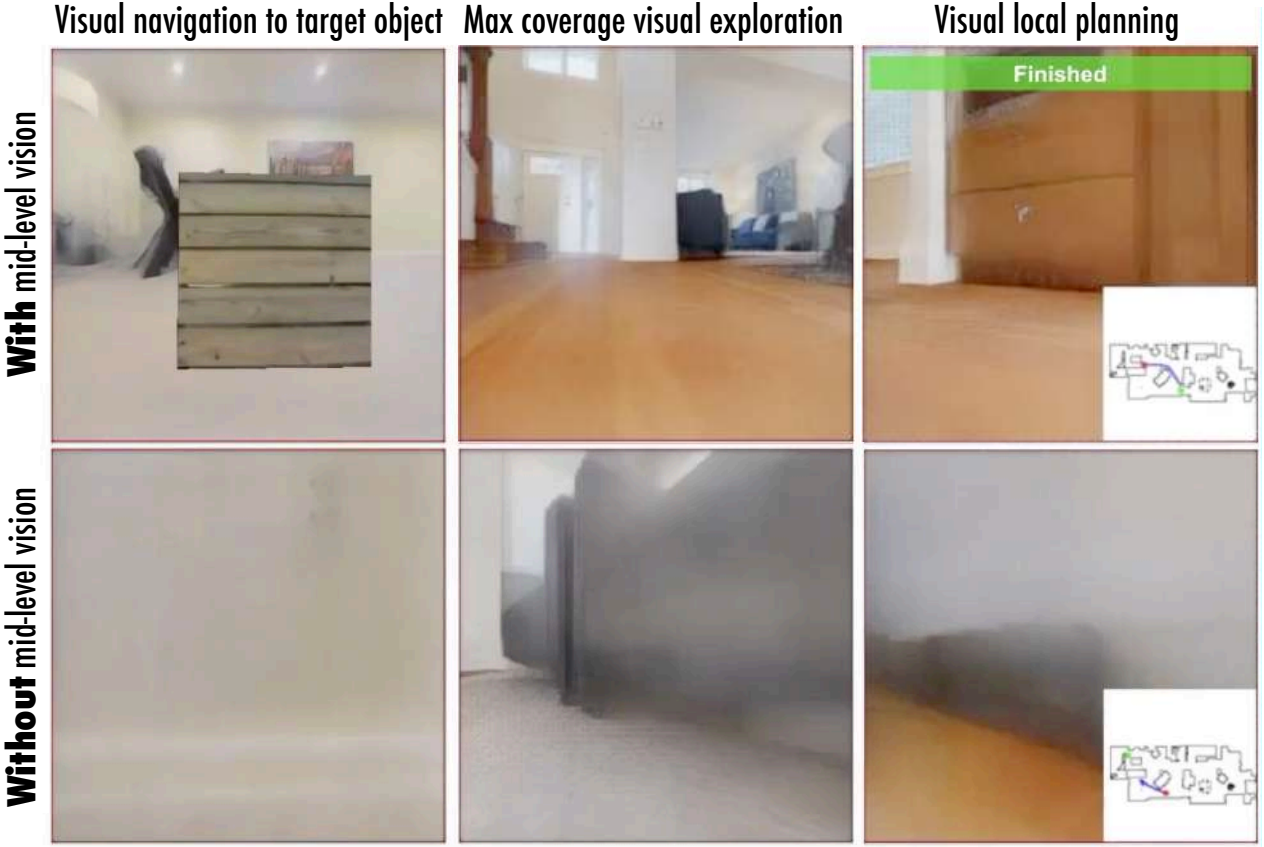
Learning with vs without perceptual priors



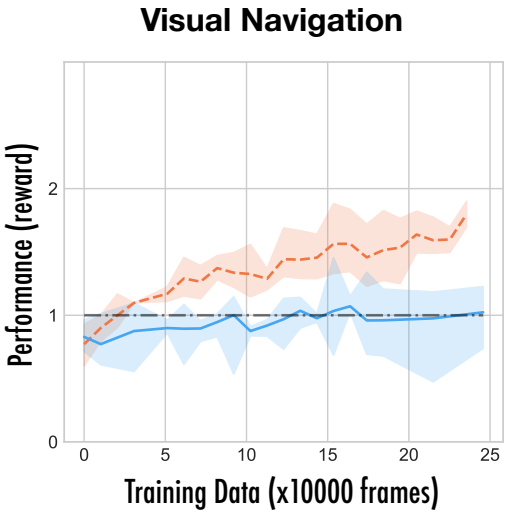
Learning with vs without perceptual priors

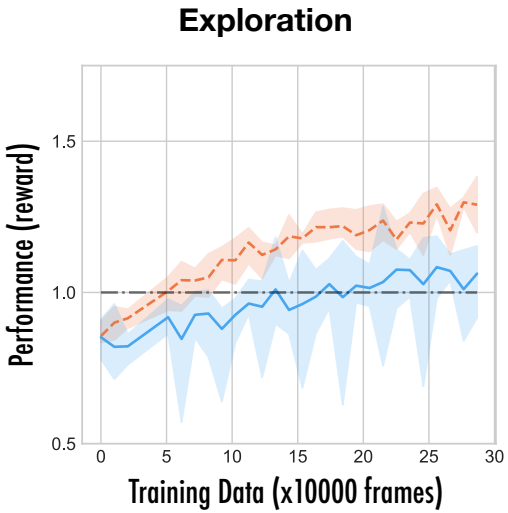


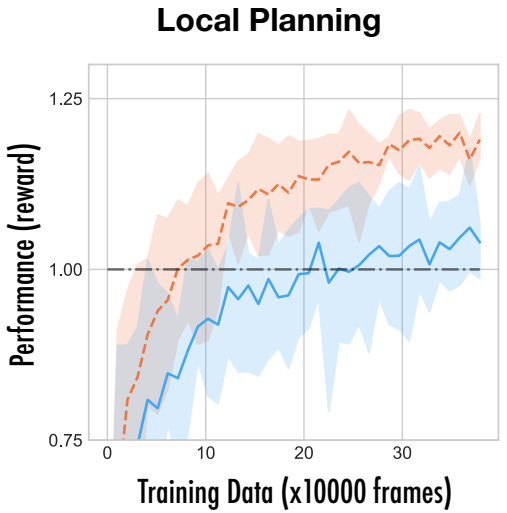
Learning with vs without perceptual priors



Learning without perpetual priors does not generalize!



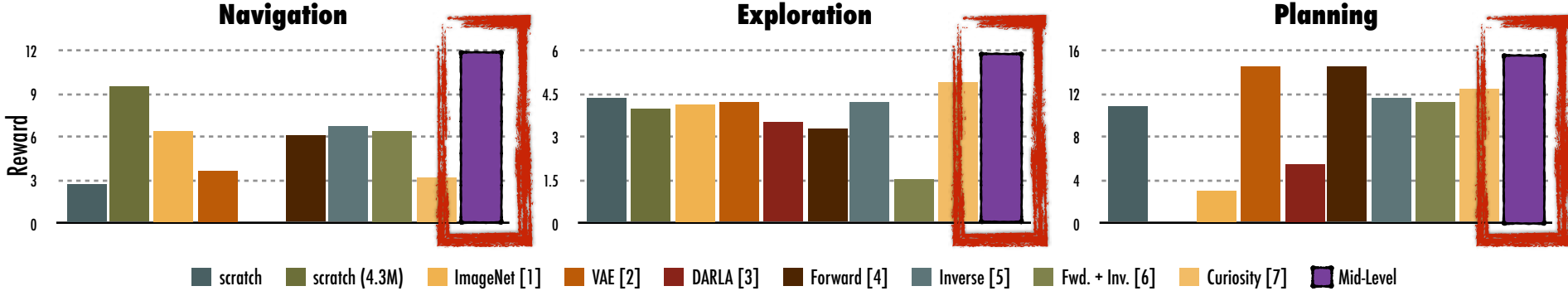




Scratch (no mid-level vision)

- Train
- Test
- Blind Agent

Comparison with SOTA feature learning



[1] A. Krizhevsky, et al. "Imagenet classification with deep convolutional neural networks" NIPS 2012
[2] S. M. A. Eslami, et al. "Neural scene representation and rendering" Science 2018
[3] I. Higgins, et al. "DARLA: Improving Zero-Shot Transfer in Reinforcement Learning" arXiv 2017
[4] J Munk, et al. "Learning state representation for deep actor-critic control" CDC 2016.
[5] E. Shelhamer, et al. "Loss is its own reward: Self-supervision for reinforcement learning" CoRR 2016.
[6] P. Agrawal, et al. "Learning to poke by poking: Experiential learning of intuitive physics. CoRR 2016 .
[7] D. Pathak, et al. "Curiosity-driven exploration by self-supervised prediction." CoRR 2017

Singleton or Set of Perception Skills?

Navigation

Feature	r	p-val
Obj. Cls.	5.91	.001
Sem. Segm.	5.87	.001
Curvature	4.75	.002
Scene Cls.	3.07	.003
2.5D Segm.	3.01	.002
2D Segm.	1.99	.003
Distance	1.74	.003
Occ. Edges	.38	.009
Vanish. Pts.	.39	.019
Reshading	.21	.021
2D Edges	.12	.006
Normals	-.50	.035
Jigsaw	-.86	.122
3D Keypts.	-1.08	.112
Layout	-1.14	.057
Autoenc.	-1.16	.043
Rand. Proj.	-2.12	.083
Blind	-3.20	.755
Pix-as-state	-4.30	.856
2D		22
In-		71
De		81

Exploration

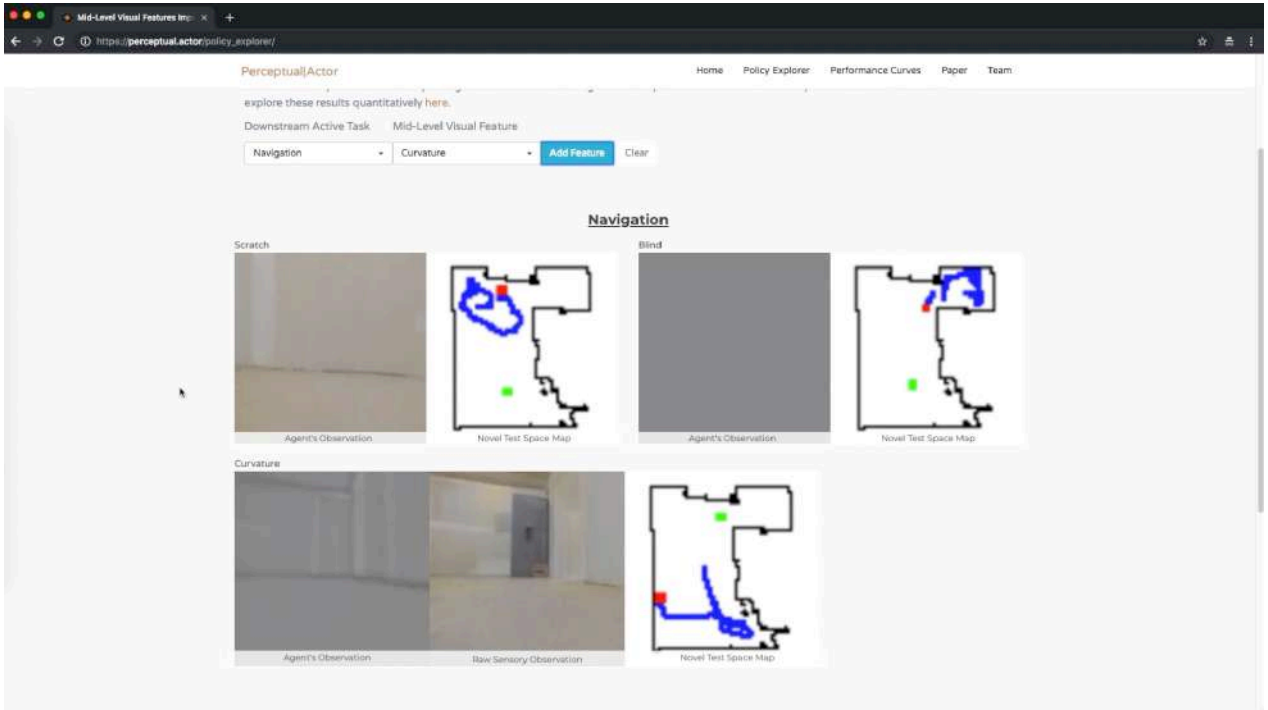
Feature	r	p-val
Distance	5.90	.015
Reshading	5.79	.003
3D Keypts.	5.27	.004
Curvature	5.12	.027
2.5D Segm.	5.60	.056
Layout	4.78	.108
2D Edges	4.87	.120
Normals	5.26	.143
Scene Cls.	4.67	.152
Obj. Cls.	4.80	.187
2D Segm.	4.47	.406
Jigsaw	4.47	.455
Rand. Proj.	4.33	.500
Vanish. Pts.	4.24	.500
Pix-as-state	4.20	.531
Blind	4.21	.545
2D Keypts.	4.21	.682
In-painting	4.30	.697
Autoenc.	4.11	.815
Ser		57
Oc		64
De		62

Planning

Feature	r	p-val
3D Keypts.	15.45	.015
Normals	15.10	.000
Curvature	14.84	.003
Distance	14.56	.001
2.5D Segm.	14.50	.001
Sem. Segm.	14.49	.000
Scene Cls.	14.20	.001
Occ. Edges	14.20	.001
Reshading	14.12	.000
Layout	14.12	.015
Obj. Cls.	13.95	.000
2D Segm.	13.86	.001
Denoising	13.54	.000
In-painting	13.28	.000
Jigsaw	13.17	.012
2D Edges	13.16	.008
Vanish. Pts.	12.14	.028
2D Keypts.	11.99	.050
Autoenc.	11.39	.155
Pix		54
Ra		92
		29

Interactive Webpage

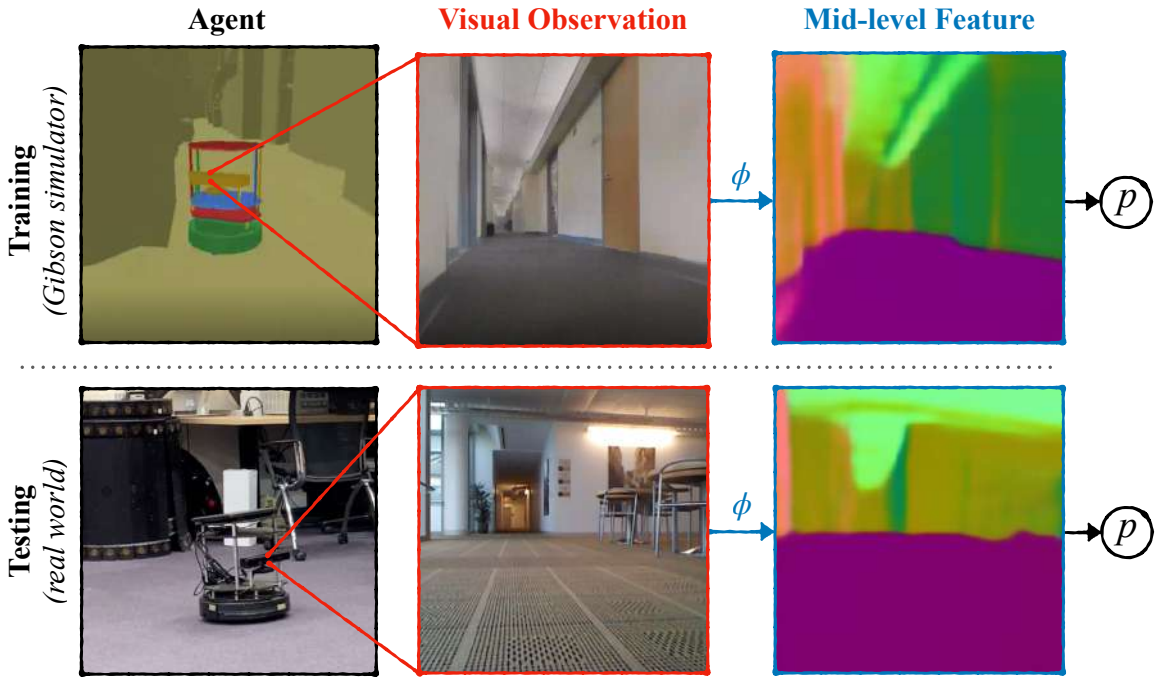
https://perceptual.actor/policy_explorer/



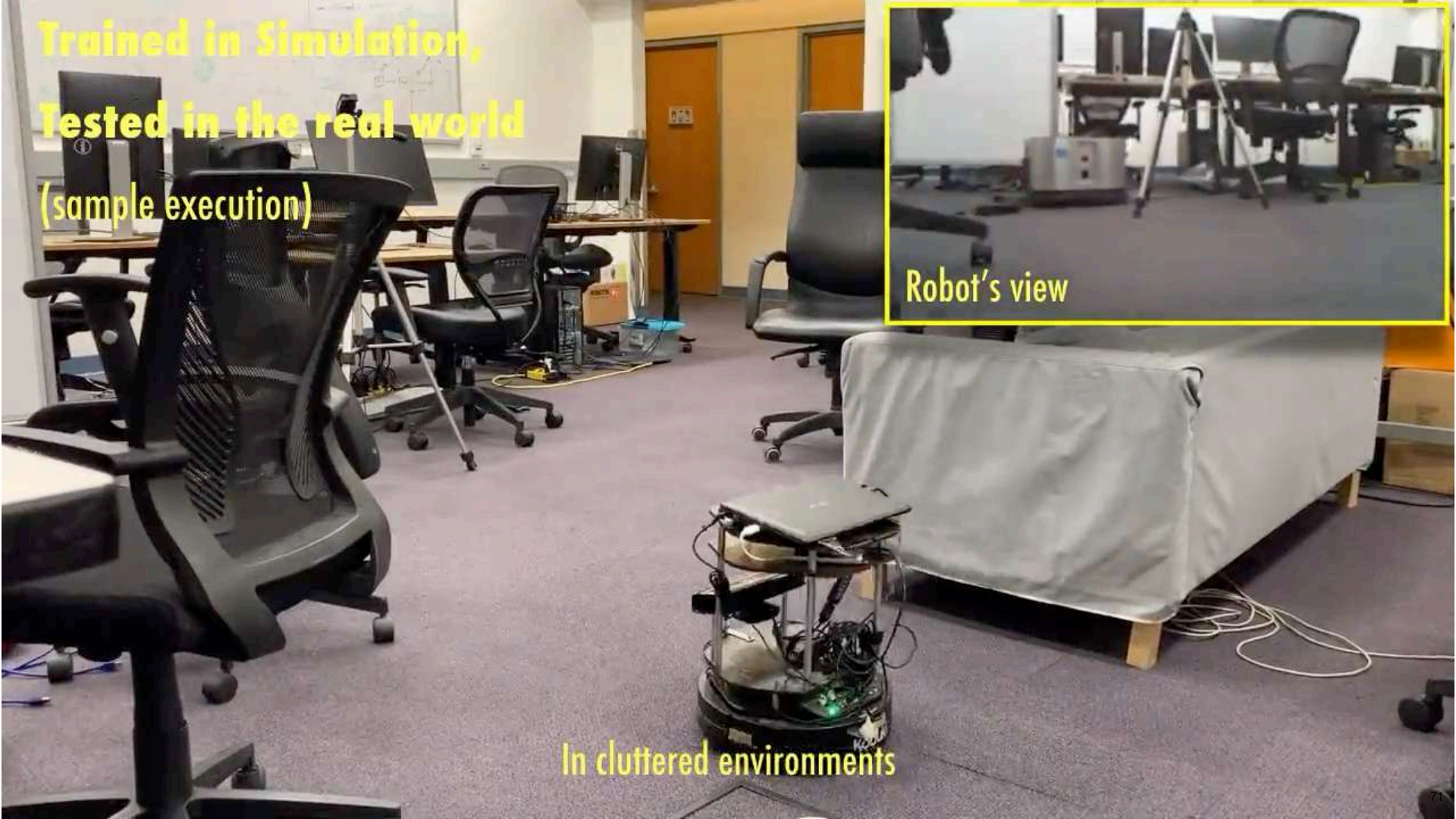
Visual Navigation Task

Visual Navigation Task

Generalization to real world

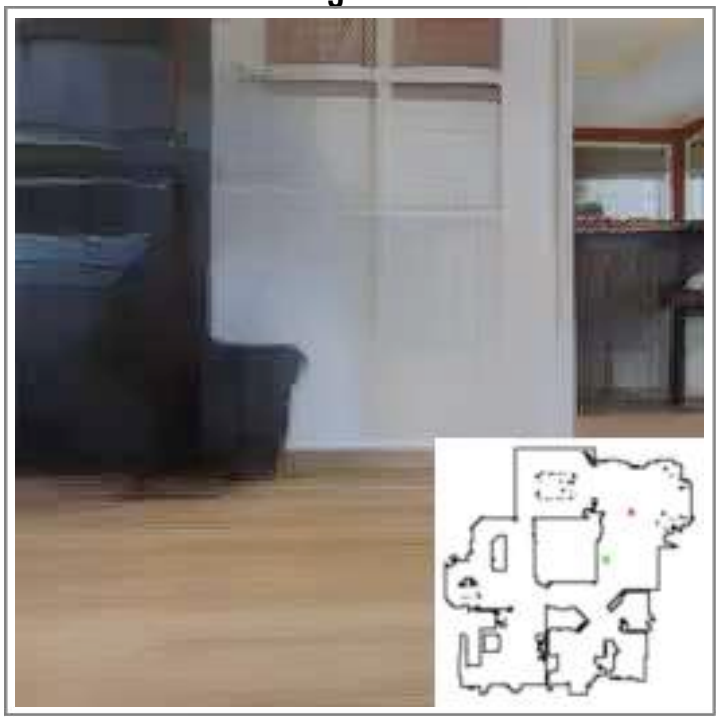


- "Robust Policies via Mid-Level Visual Representations: An Experimental Study in Manipulation and Navigation". Chen, Sax, Pinto, Lewis, Armeni, Savarese, Zamir, Malik. CoRL20.


Trained in Simulation,
Tested in the real world
(sample execution)
Robot's view
In cluttered environments

Generalization to Real-World

Training in Gibson



(task: find the orange target and navigate to it)

- "Robust Policies via Mid-Level Visual Representations: An Experimental Study in Manipulation and Navigation". Chen, Sax, Pinto, Lewis, Armeni, Savarese, Zamir, Malik. CoRL20.

Generalization to Real-World

Training in Gibson

(task: find the orange target and navigate to it)

- "Robust Policies via Mid-Level Visual Representations: An Experimental Study in Manipulation and Navigation". Chen, Sax, Pinto, Lewis, Armeni, Savarese, Zamir, Malik. CoRL20.

Generalization to Real-World

Training in Gibson

Testing on Real Robots

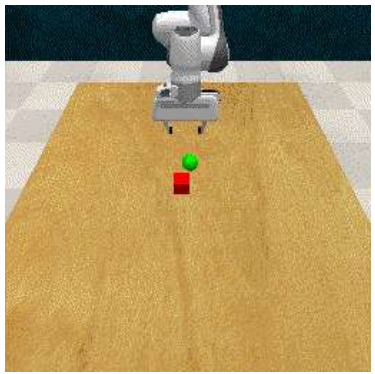
(task: find the

- "Robust Policies via

Mid-Level Vision for Manipulation

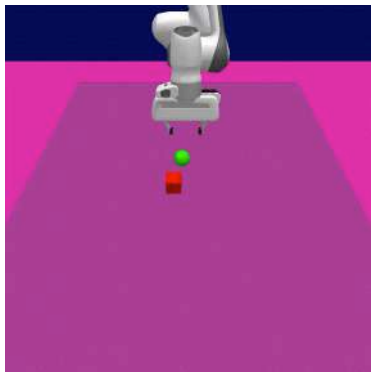
Train

Pick Red Object and **Place** at Green Sphere

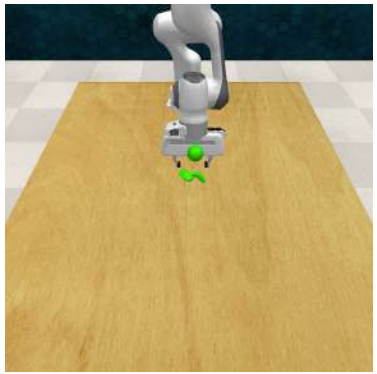


Test Environment (new texture, new object, etc)

With Mid-Level Representations

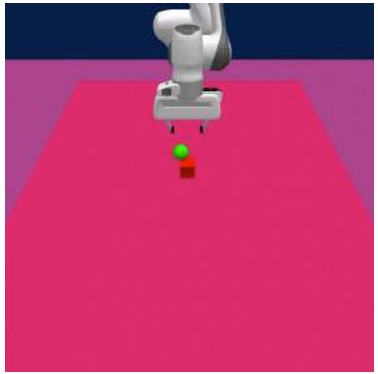


Novel Texture

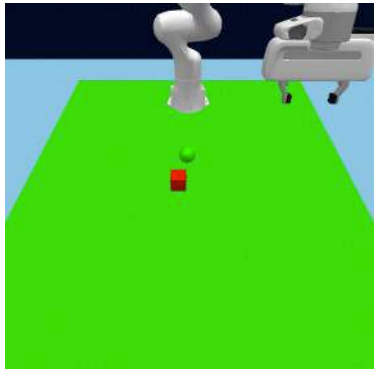


Novel Object

Without Mid-Level Representations



Using **Domain Randomization**

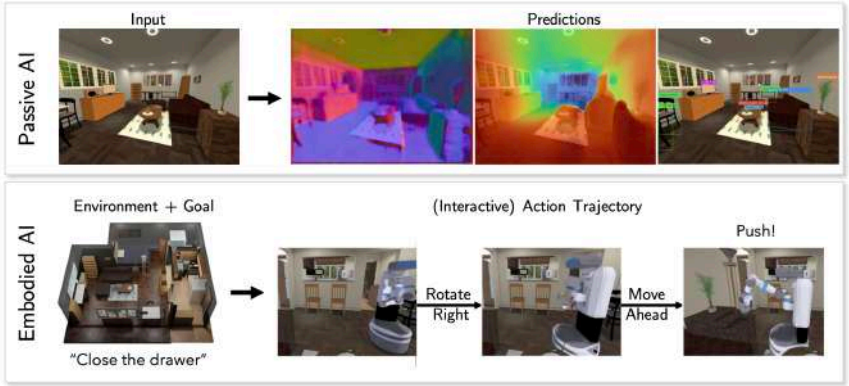


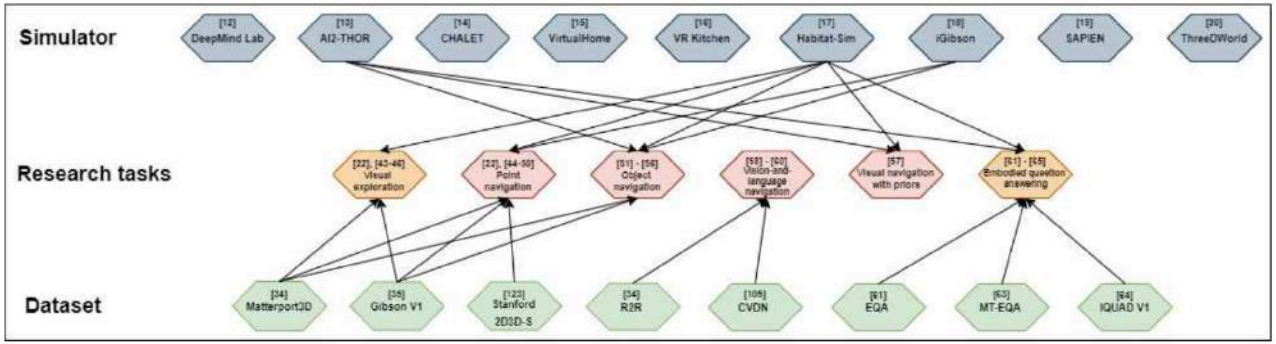
- "Robust Policies via Mid-Level Visual Representations: An Experimental Study in Manipulation and Navigation". Chen, Sax, Pinto, Lewis, Armeni, Savarese, Zamir, Malik. CoRL20.

Standardized embodied vision efforts (as of ’24/’25)

Common Tasks (2023)

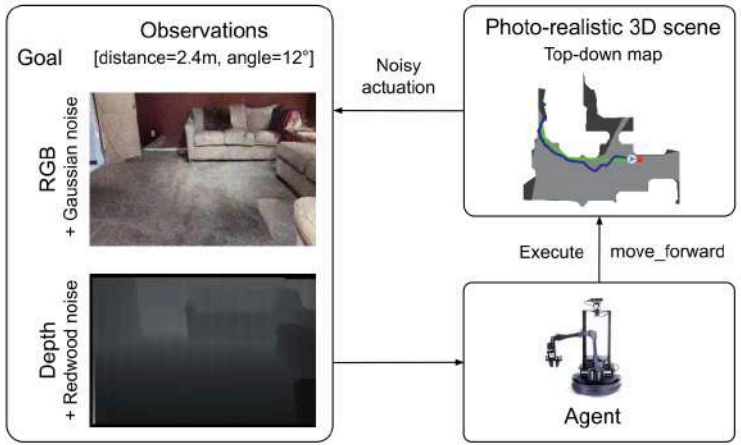
- (1) visual navigation
- (2) rearrangement
- (3) embodied vision-and-language





Retrospectives on the Embodied AI Workshop, Deitke et al., 2022

PointNav

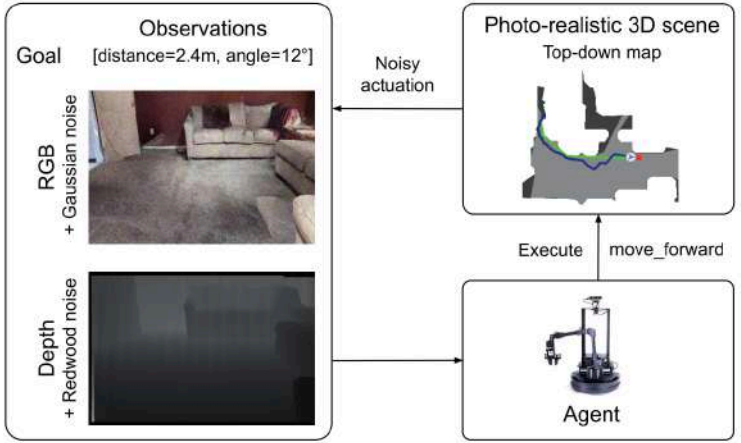


Retrospectives on the Embodied AI Workshop, Deitke et al., 2022

PointNav++

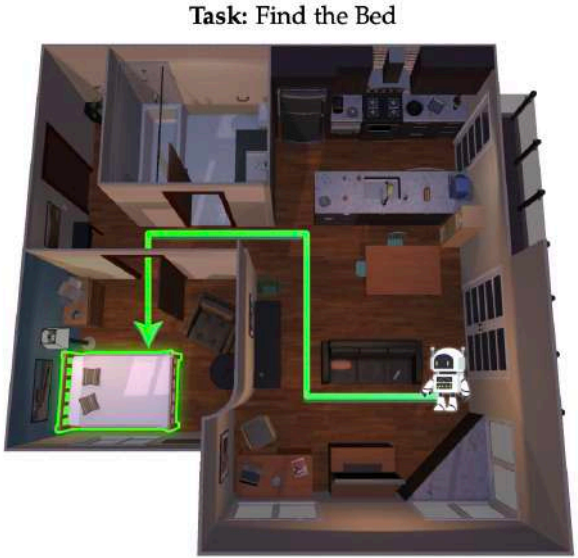
(a) Interactive Navigation

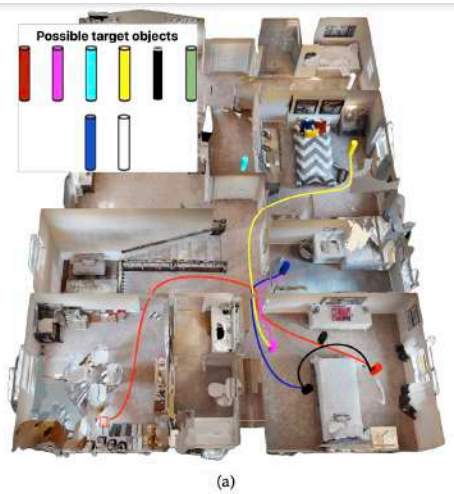
(b) Social Navigation



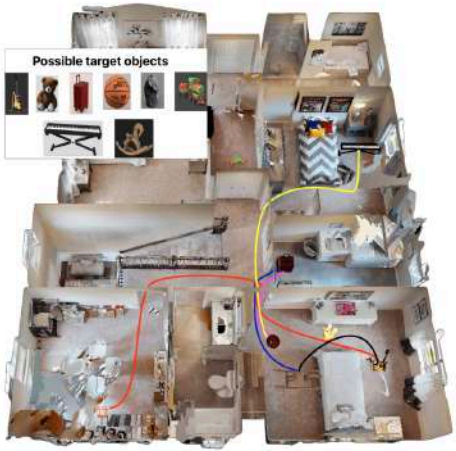
Retrospectives on the Embodied AI Workshop, Deitke et al., 2022

ObjectNav





(a)



(b)

Retrospectives on the Embodied AI Workshop, Deitke et al., 2022

Nav++

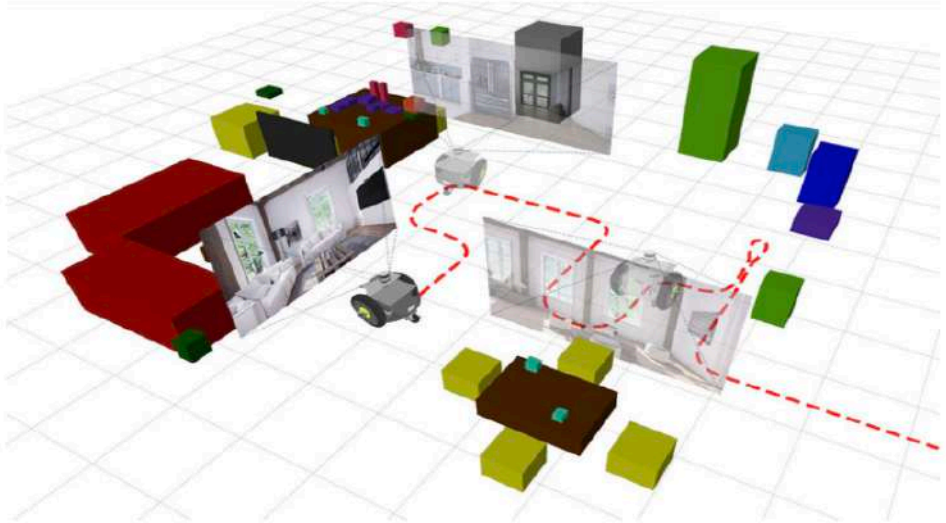
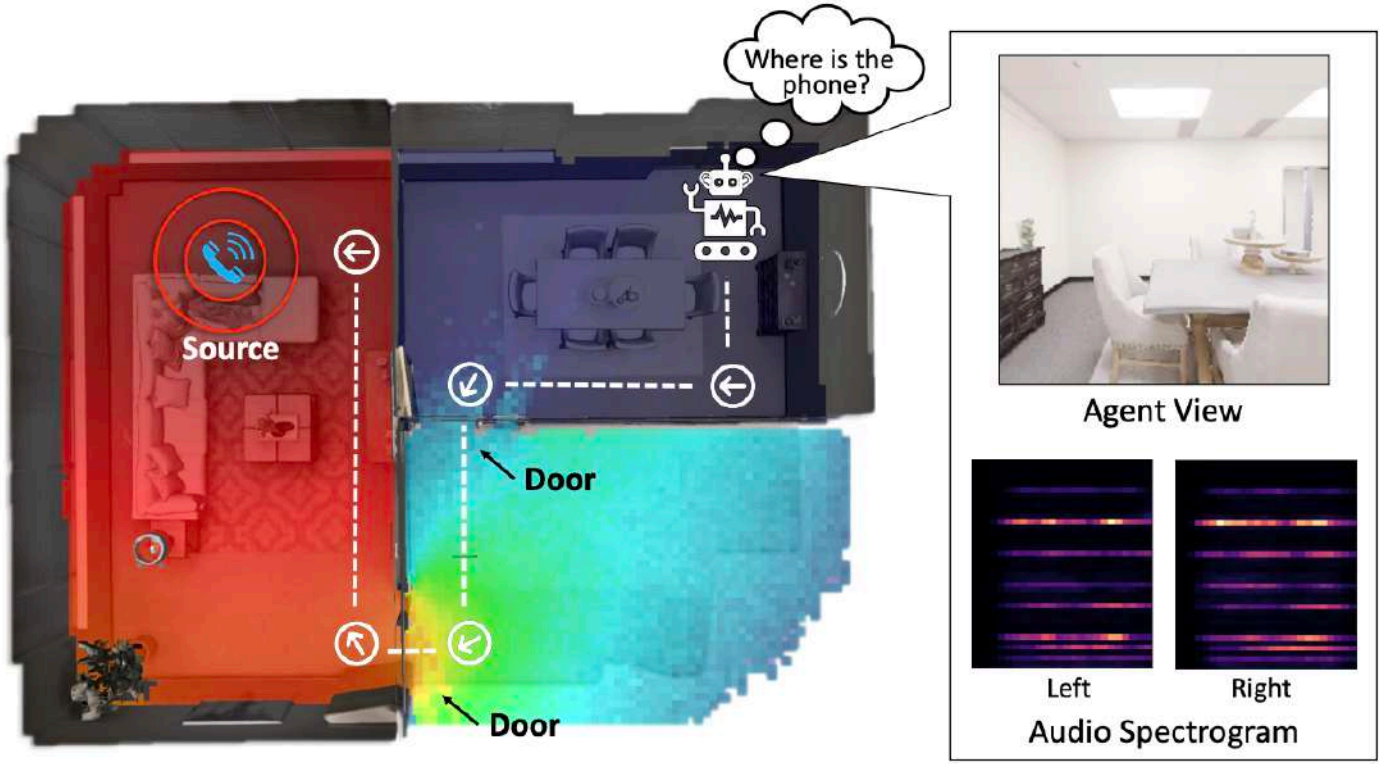


Figure 7. In the *RVSU Semantic SLAM* task, an autonomous agent explores environment to create a semantic 3D cuboid map of objects.

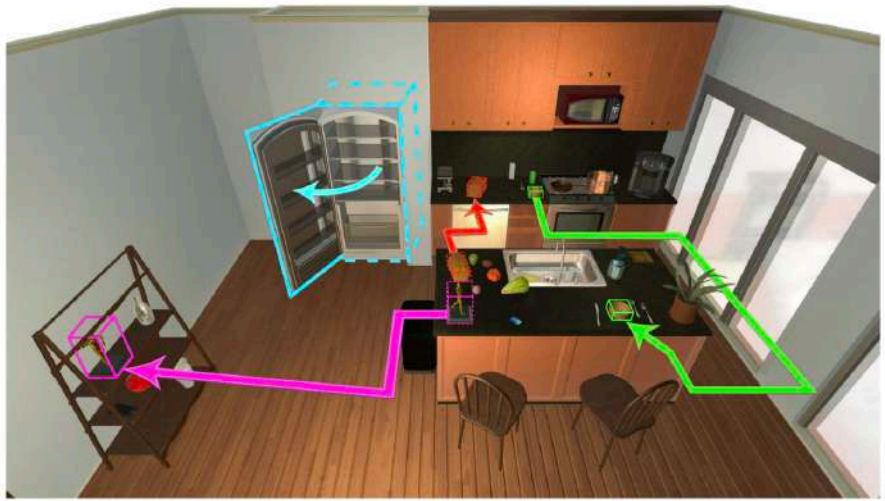
Retrospectives on the Embodied AI Workshop, Deitke et al., 2022

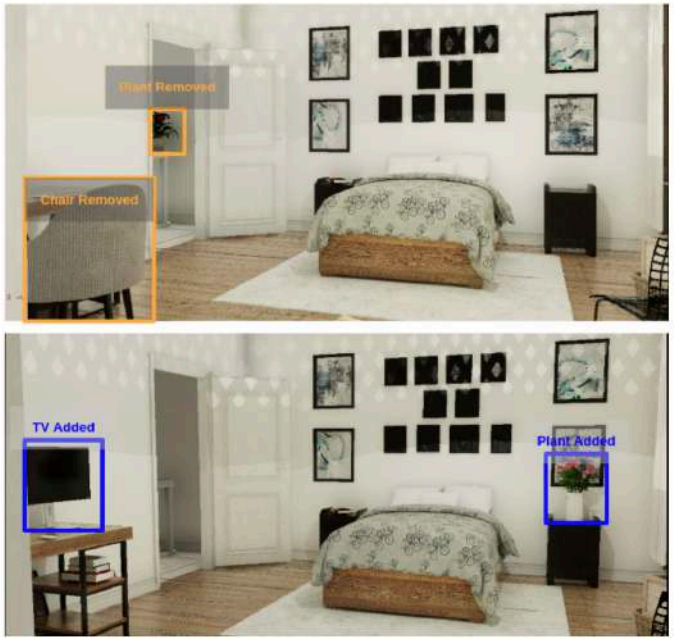
Nav++ (multimodal)



Retrospectives on the Embodied AI Workshop, Deitke et al., 2022

Rearrange





Retrospectives on the Embodied AI Workshop, Deitke et al., 2022

Embodied vision-and-language

You are in a bedroom. Turn around to the left until you see a door leading out into a hallway, go through it. Hang a right and walk between the island and the couch on your left. When you are between the second and third chairs for the island stop.

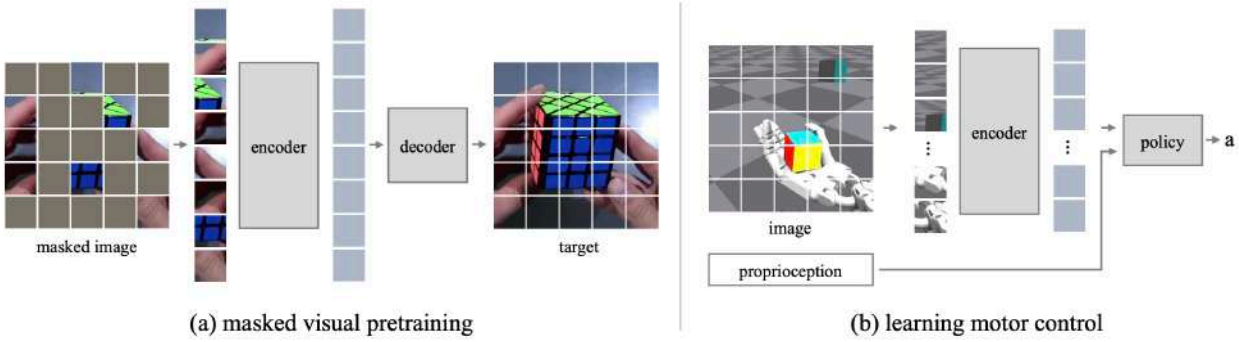
Retrospectives on the Embodied AI Workshop, Deitke et al., 2022

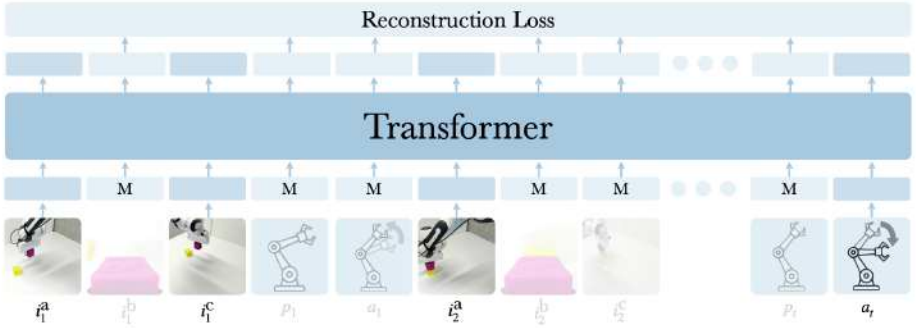
Empirical lessons

		Best End-to-end			Best Modular		
Challenge	**Simulator**	**Method**	**Success**	**Rank**	**Method**	**Success**	**Rank**
ObjectNav	Habitat	Habitat-Web	60	2	Stretch	60	1
Audio-Visual Navigation	SoundSpaces	Freiburg Sound	73	2	colab_buaa	78	1
Multi-ON	Habitat	-	-	-	exp_map	39	1
Navigation Instruction Following	VLN-RxR	CMA Baseline	13.93	10	Reborn	45.82	1
Interactive Instruction Following	AI2-THOR	APM	15.43	14	EPA	36.07	1
Rearrangement	AI2-THOR	ResNet18 + ANM	0.5	6	TIDEE	28.94	1

Retrospectives on the Embodied AI Workshop, Deitke et al., 2022

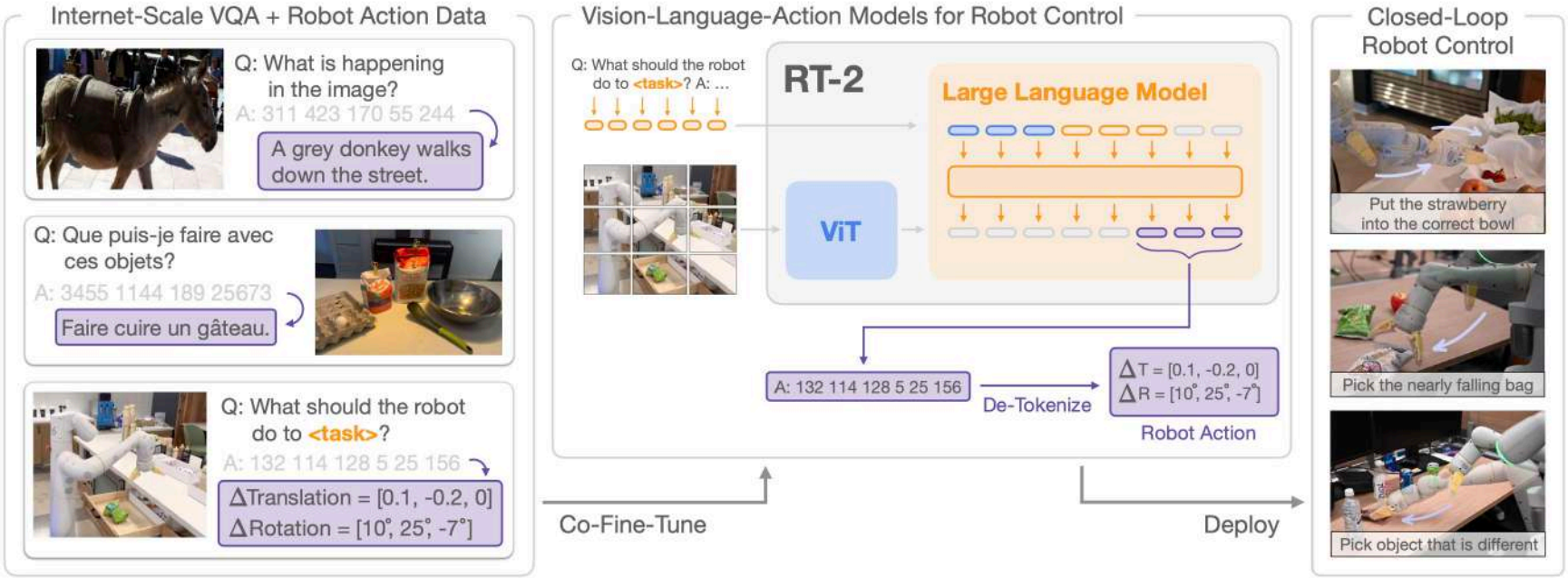
Multi-modal learning → Motor Control





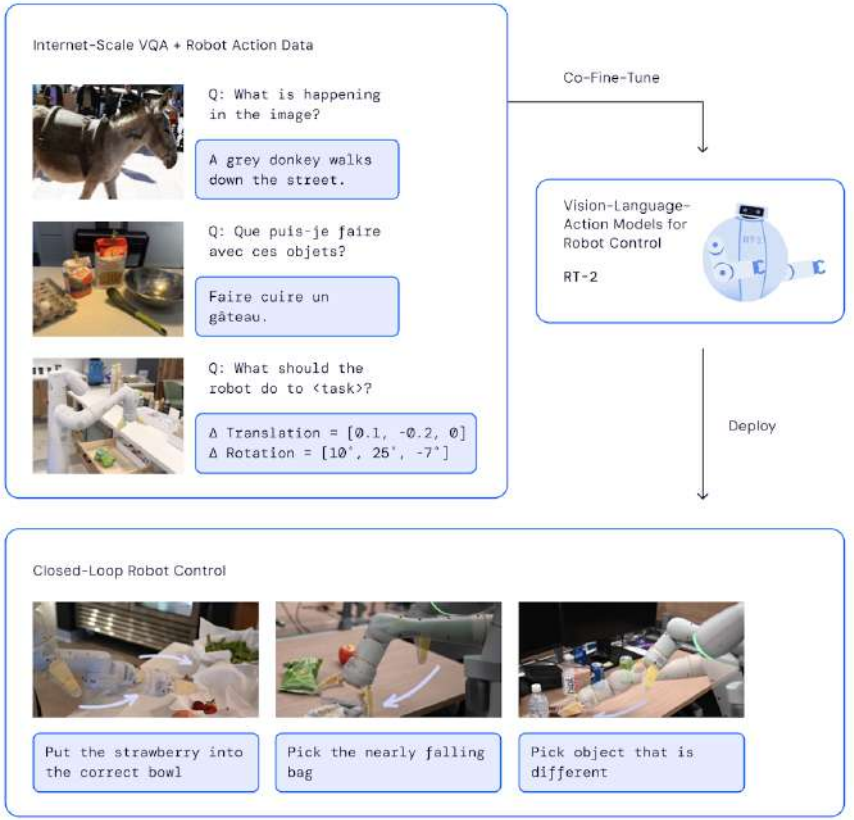
Robot Learning with Sensorimotor Pre-training, Radosavovic, Shi, Fu, Goldberg, Darrell, Malik. 2023
Real-World Robot Learning with Masked Visual Pre-training, Radosavovic, Xiao, James, Abbeel, Malik, Darrell. CoRL 2022
Masked Visual Pre-training for Motor Control, Xiao, Radosavovic, Darrell, Malik. ArXiv 2022
MultiMAE: Multi-Modal Multi-Task Masked Autoencoders, Bachmann, Mizrahi, Atanov, Zamir. ECCV 2022

Multi-modal learning → Motor Control



RT-2: Vision-Language-Action Models Transfer Web Knowledge to Robotic Control, 2023.
PaLM-E: An Embodied Multimodal Language Model, 2023.

LLMs in robotics pipelines



- RT-2: Vision-Language-Action Models Transfer Web Knowledge to Robotic Control, Google, 2023.

Questions?

https://vilab.epfl.ch/